SINGAPORE POLYTECHNIC

School of Mathematics & Science

Business Statistics

EP0607 / MS0140 / MS1100 / MS7124

AY 20/21 Semester 2

SCHOOL OF MATHEMATICS & SCIENCE

BUSINESS STATISTICS

Module Code: EP0607 / MS0140 / MS1100 / MS7124

Instructional Hours

Lecture:	15 hours
Tutorial:	45 hours
Total:	60 hours
Credit Units:	4

Module Aims

Provides foundation for students to be equipped with quantitative skills, understanding of basic statistical concepts and their relevance in business. It is designed to train students with the statistical research skills from data analysis through manual means and software, data representation and interpretation that will allow them to make informed decisions. The statistical problem-solving process is taught as a method in addressing business-related statistical problems. Topics covered include descriptive statistics, probability distributions, sampling, estimation, hypothesis testing, analysis of variance, and linear regression.

Recommended Text

Triola, M.F., 2008. Essentials of Statistics. 3rd ed. Pearson

References

Freund, J.E. and Perles, B.M, 2007. Modern Elementary Statistics. 12th ed. Pearson.

Assessment

Component	Weight	Details
MST (Mid-Semester Test)	25 %	Minitab-assisted paper test
EST (End-of-Semester Test)	30 %	Minitab-assisted paper test
CA (Continuous Assessment)	45 %	Weekly Class Quizzes Weekly in-class participation & pre-class tasks Poster project

Lecturer

Name	Room	Office	Email

Module Teaching Schedule

Academic Year 2020/2021 Semester 2

Business Statistics

Week	Chapter	Note
1	0. Introduction to Statistics 1. Descriptive Statistics	
2	1. Descriptive Statistics	
3	2. Probability Distributions	
4	2. Probability Distributions	
5	3a. Sampling Distribution of the Sample Mean 3b. Estimating Population Mean	
6	4. Hypothesis Testing of Mean	
7	Revision	
8	**Mid-Semester Test (MST)**	
9	Vacation	
10		
11		
12	5. Concepts of Hypothesis Testing	
13	6. Hypothesis Testing of Two Population Means	
14	7. ANOVA	
15	8. Chi-Square Testing	
16	9. Simple Linear Regression	Project Submission
17	Revision	Project Submission PH: CNY – 12/2 Fri
18	**End-Semester Test (EST)**	

CHAPTER 0
INTRODUCTION TO STATISTICS

Learning Objective:

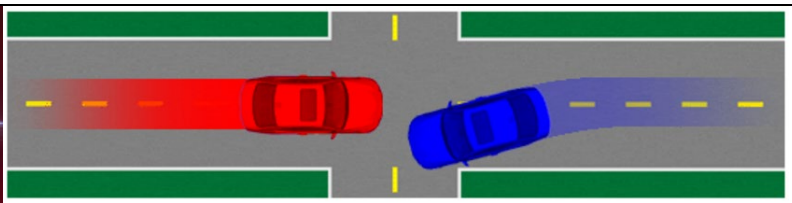
1. *To appreciate statistics as a useful tool to learn more about events around us.*

---

1. Introduction

In general terms, what a working engineer does is to design, build, operate, and/or improve physical systems and products. Civil engineers build highways, waterworks and large buildings; chemical engineers design and operate systems related to fertilizers to fuels, aeronautical and aerospace engineers design and improve aircraft systems, industrial engineers design and operate manufacturing facilities, etc.

As technology advances and new systems and products are encountered, engineers are often faced with questions and situations where their existing knowledge and experience may offer little or no help. Hence, what do engineers do?



Do autopilot self-driving cars generally have a significantly better (lower) reaction time compared to human beings in emergency situations?

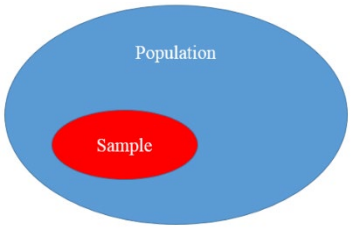
It is necessary then for engineers to ask the correct question, collect unbiased data, analyze data rigorously, and then interpret the results to help them understand how the new system or the new product works. If data are collected haphazardly or analyzed poorly, then valuable time and resources are wasted. Worse, sometimes erroneous conclusions made may lead to bad decisions which could cost lives.


Retrieved from: https://www.nasa.gov/multimedia/imagegallery/image_gallery_2437.html

The NASA family lost seven of its own on the morning of Jan. 28, 1986, when a booster engine failed, causing the shuttle Challenger to break apart just 73 seconds after launch.

2. Statistics

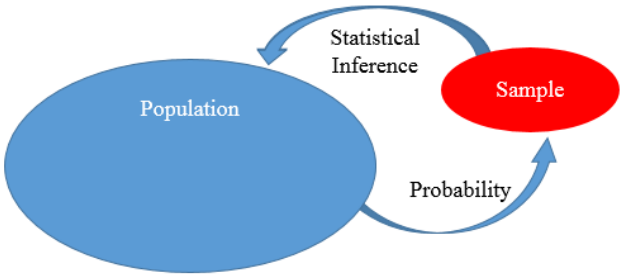
So what is statistics? It is a scientific way of thinking that helps us understand more about the phenomena that we encounter in our workplace or in our daily lives – performance of autopilot cars, likelihood of space shuttle exploding, the quality of potato chips, and many others.



To better understand what statistics is, let us first look at the concept of sample and population.
The **population**, whether it is people, grapes or marbles, is the entire group we want information about.
A **sample** is part of the whole. In other words, sample is a subset of population.

If we can access every member of the population to collect data, this will be a **census**.
It seems like a census would be a straightforward way to get the most accurate, thorough information, but taking an accurate census is not nearly as easy as one might think. It could be due to ignorance about size of population, or sheer size of population, or inaccessibility of its members.

Thus, the sample is the part we actually gather data from. Since the sample is only a part of the population, we can study it more extensively than we can all the members of the population. Then we can use the sample data to draw conclusions about the entire population (statistical inference). Conversely, knowledge about the population would tell us how likely we are to get the sample we obtained (probability).



For these conclusions to be valid though, the sample must be representative of the population. To make sure it is, statisticians rely on what is called simple **random sampling**. This means that the sample is chosen in such a way that each member has an equal chance of being selected. This helps eliminate bias in the study design and the conclusion drawn based on the sample could be generalized to the entire population with high level of confidence.

Retrieved from: "Census and Sampling: Against All Odds—Inside Statistics," director. Films Media Group, 2013, http://fod.infobase.com/portalplaylists.aspx?wid=151497&xtid=111535.

Frito Lay potato chip fans count on consistent appearance and taste from their favorite brand, and sampling is one way the company meets those expectations.

3. Outline of Content

The aim of statistics in this module is to provide you with a strong foundation in the scientific thinking of statistics that would be useful in solving engineering problems involving statistical data.

Hence in Chapter 1, you will be introduced to the statistical problem-solving process:

You will also learn important skills in collecting, managing, summarizing and presenting data in an informative manner through tables and charts using a statistical software called **Minitab**.

In Chapter 2, you will be taught the basic concept of probability as a measure of likelihood of an event happening. Then, the concept of probability will be extended to probability distributions where real-life scientific case study such as the explosion of space shuttle Challenger will be discussed. The idea of "rare-event" will also be introduced.

In Chapter 3A, the very important concept of sampling distribution and the Central Limit Theorem will be presented. You will look at a case study involving claims about Subway selling less than 12-inch foot-long subway sandwiches. The idea of "rare-event" would be further discussed using more data-driven examples, motivating the concept of "P-value" in Chapter 4.

In Chapter 3B, you will learn how to estimate the value of an unknown population parameter, specifically the population mean. Instead of using only one value to estimate an unknown population parameter, you will be taught the concept of confidence interval to construct an interval of estimates for the unknown population parameter. In Chapter 4, you will learn how to identify statistical hypotheses to be tested and the powerful hypothetical probabilistic reasoning in drawing conclusions about a population based on just a sample of data. The concept of P-value which is ubiquitous in decision-making will be presented formally.

In Chapter 5, you will learn about the relationship between test statistic and P-value, what are critical values, and how to use them to make a decision about the statistical hypotheses. You will also learn about errors that may occur in hypothesis testing.

In Chapter 6, you will know how to make use of appropriate tests to compare and make inference about the means from two populations. We will take this one step further in Chapter 7 where we use Analysis of Variance to compare three or more population means.

In Chapter 8, you will be taught how to analyze qualitative variables using the Chi-Square test. This test allows us to find out if one variable has the same distribution across two or more groups, or if there is an association between two variables.

Finally in Chapter 9, you will learn how to model a linear relationship between an explanatory variable and a response variable, and use a linear model to make prediction.

CHAPTER 1
DESCRIPTIVE STATISTICS

Learning Objectives :

1. *Understand statistics as a methodology that is concerned with formulating question, collecting data, analyzing data and interpreting results.*
2. *Use basic terminology in statistics such as random variable, population and sample.*
3. *Distinguish between the different types of data, such as qualitative and quantitative.*
4. *Construct various graphical displays of the data, and provide basic interpretations.*
5. *Compute numerical summaries of the data, and provide basic interpretations.*
6. *Analyse strength of relationship using scatter plot and correlation coefficient*

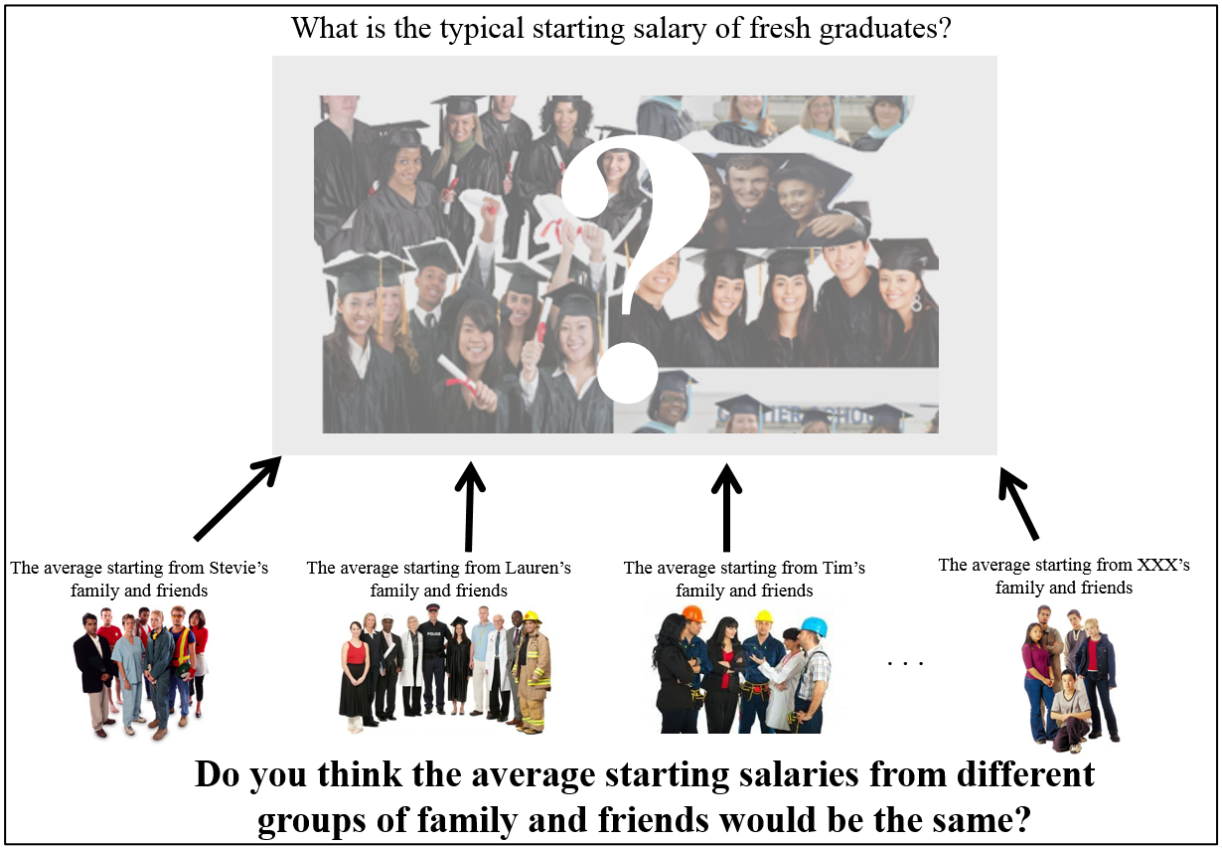
Content

1. Statistical Problem-Solving Process

Suppose we are interested to find out the *typical* starting salary of a fresh graduate, how would we go about finding an appropriate answer?

We could ask relatives and friends to share with us their starting salaries and then, perhaps we would take an average of their salaries. This average number, we presume, is possibly a typical starting salary based on the data we have collected.

Suppose now there are ten other people who are interested in the same question and they took samples of their own relatives and friends (for convenience, assume that none of these ten people are related and that they do not share the same relatives and friends), would you expect all of these ten people to arrive at the same typical starting salary as ours earlier?

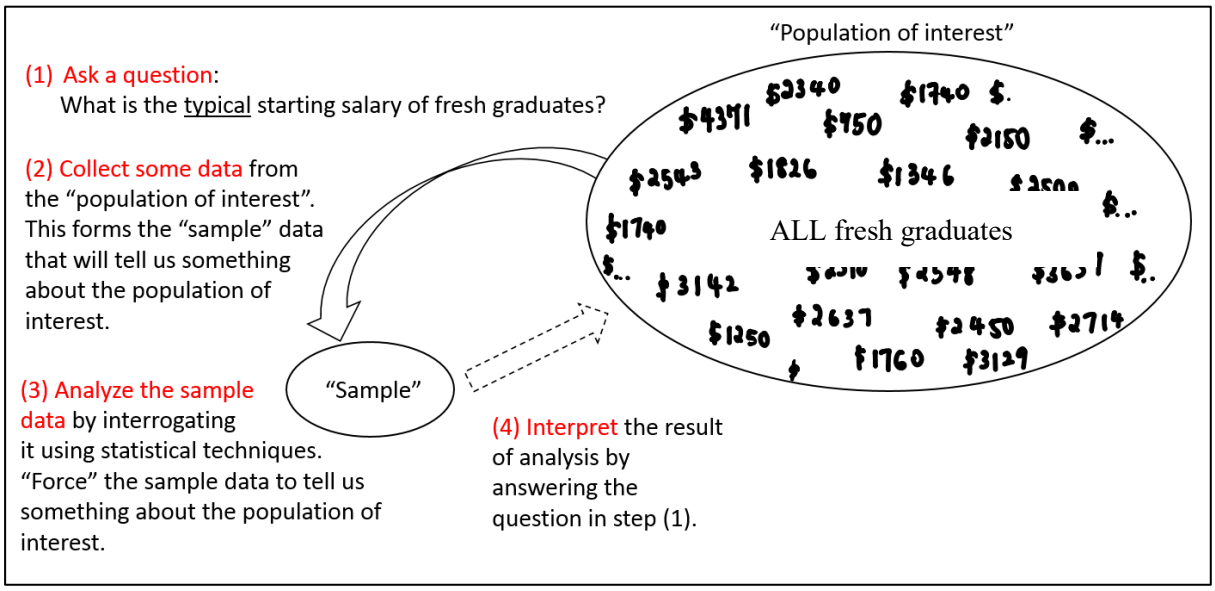


Well, it is highly likely the average salaries from different groups of family and friends would be different.

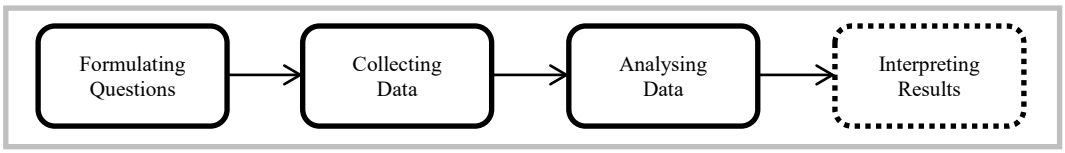
In studies where there are **variability**, we need a tool to help us capture this variability.
How do we capture the *variability* in different sample data sets and use it to make more sense of the data sets?

One tool we can use to study data – the typical value, the variability of data, and more – is **statistics**.

Essentially, this is what we are doing:



Steps (1) to (4) outlined the **statistical problem-solving process**, summarized here:

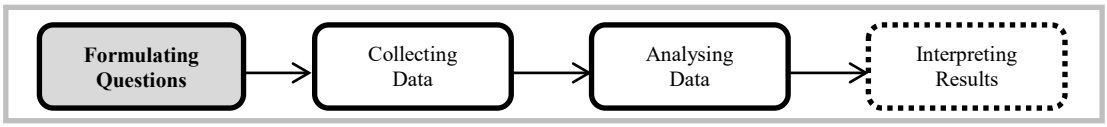


You may have learnt a lot of step (3) in school, so it is hoped that this chapter will value-add to your learning by teaching you the holistic statistical problem-solving process which always begins with a **question** of interest, then **collect** some data to help answer the question, **analyze** the data by using statistical techniques, then **interpret** the results to answer the question! ☺ Thus, the focus of this chapter is to take you through the statistical problem-solving process of steps (1) to (4) through a case study – Prestige Mall.

Remember, the main aim of this chapter is to give you an opportunity to experience statistics in a more holistic way, hence every step of the statistical problem-solving process is as important as the other steps. And to remind you of the step which you are at, the process can be found at the top of every page with the current step highlighted.

Exercise: Can you identify the population and sample in the scenarios below?

Scenario	Population	Sample
A new filtration system has been installed in the water systems of a small city. The amount of impurities (in parts per million) remaining in the filtered water is recorded over a 30-day period.	Filtered water of the city.	
To serve customers better by cutting queueing time at counters during peak period, ABC Bank recorded the queueing times (in minutes) of 20 customers.		The 20 customers whose queueing times were recorded.



2. Formulating Questions

Case Study: Prestige Mall

What businesses have you here?

One of the shops in this shopping centre, *Prestige Mall*, has become vacant. You were tasked by your boss to suggest a possible tenant to bring into the Mall.

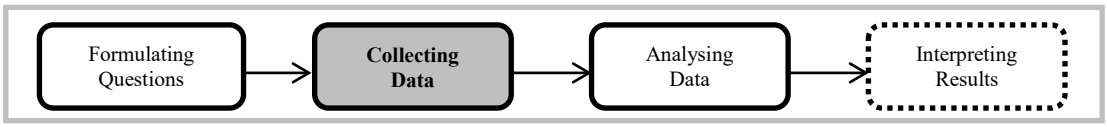
Some decisions that we make may be based on personal judgments but some may not. In this case, a proposal with support from data collected is obviously more convincing as compared to a proposal without such a support or a proposal supported with personal judgments alone. What are some of the information that you would include in your proposal, as support to the kind of business (and thus the potential tenant) that you would like to recommend to your boss?

Well, for instance, you may like to have an initial "feel" of the profile of customers who patronize this mall:

- What kind of job sectors are they from?
- How many times do they frequent Prestige Mall in a month?
- How much do they typically spend in Prestige Mall?
- What is their age profile?
- What is the proportion of male customers of Prestige Mall?
- What is their average monthly household income?

Think-out-loud…
So, what are some other questions that you might ask?

The questions listed above would generate valuable information in helping you to decide on the potential tenant that you would like to recommend to your boss.



3. Collecting Data

3.1 Sample Data

The previous section introduces a case study, specifically what kind of business opportunities there are in Prestige Mall. The rest of this chapter would base its contents and discussions on the given case study, guided by the statistical problem-solving process, as indicated by the top of each page.

We have asked a few questions in the previous section and to answer those questions, we would need to collect data. There are many possible ways to collect data, such as from Prestige Mall's database, provided there is one. However, if interrogating databases is not a viable option, then another possible method is to conduct a survey. It is not uncommon to see people filling up survey forms in malls, of course after proper permission is sought. A possible survey form could look like this:

Prestige Mall Customer Survey

Dear Valued Customer,

Thank you for taking part in this survey. Your feedback will be valuable in helping us to enhance your shopping experience here in Prestige Mall. This survey will take approximately 5 minutes. All information shared with us will remain private and confidential.

---

(1) Which job sector are you in? □ IT/Engineering □ Business/Finance □ Others

(2) What is your gender? □ Male □ Female

(3) What is your age? _____

(4) How many times did you visit Prestige Mall in the past month? _____

(5) How much is your monthly household income? _____

(6) Approximately how much do you spend at Prestige Mall a month? _____

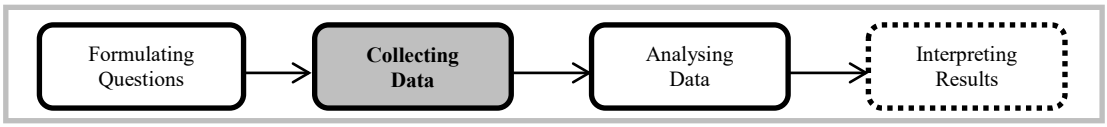
Kindly drop this survey form at the information counter and receive a token of appreciation.

Thank You!

Think-out-loud…
Why not **all** the customers?

In order to collect data, a selected group of customers of Prestige Mall is to be chosen to respond to the survey. In statistics, we are concerned with randomness and representativeness – how do we know that we have not been biased in selecting the respondents for the survey?

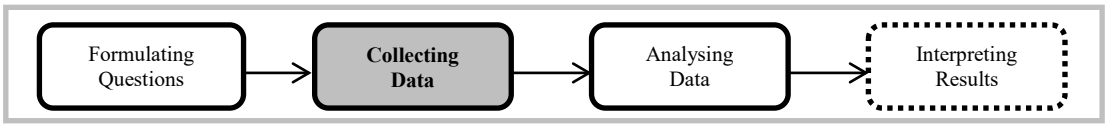
There are methods of sampling in statistics such as simple random sampling, systematic sampling, stratified sampling, and many more; but it is beyond the scope of this chapter to discuss sampling methods further. Hence we would make an assumption that all the customers who eventually are involved in the survey above are randomly selected (hence, not biased).



By the end of the survey period, suppose you have collected feedback from 200 customers. You then entered the data into an Excel spreadsheet as shown below; sorted first by gender, then by age.

No.	Job sector	Age	No. of visits per month	Gender	Household income (in $)	Amount spent per month (in $)
1	Bus/Fin	18	6	Female	8852.32	441.73
2	Bus/Fin	19	3	Female	7889.24	663.63
3	Bus/Fin	22	1	Female	6901.79	489.26
4	Bus/Fin	22	6	Female	8566.96	412.19
5	Bus/Fin	23	6	Female	7144.45	188.59
6	Bus/Fin	24	6	Female	8032.08	253.59
7	IT/Eng	25	5	Female	9405.84	421.35
8	Bus/Fin	25	3	Female	7694.62	538.59
9	Bus/Fin	25	3	Female	8727.36	489.49
10	Bus/Fin	25	5	Female	7723.2	514.52
11	Bus/Fin	26	3	Female	9598.72	528.6
12	Others	27	3	Female	6943.93	617.12
13	Others	27	4	Female	7628.78	387.54
14	Bus/Fin	28	6	Female	6584.36	607.96
15	Bus/Fin	28	4	Female	8018.32	395.89
16	Bus/Fin	28	2	Female	8532.96	445.36
17	Bus/Fin	29	4	Female	8000	430.04
18	Bus/Fin	30	6	Female	7940.7	529.17
19	IT/Eng	30	4	Female	7041.06	485.61
20	IT/Eng	30	4	Female	8008.96	549.98
21	IT/Eng	30	5	Female	9364.32	634.15
22	IT/Eng	31	4	Female	7801.48	249.33
23	Bus/Fin	32	6	Female	7686.12	393.87
24	Others	32	2	Female	6260.83	546.15
25	IT/Eng	33	2	Female	8072.96	462.99
26	Bus/Fin	34	3	Female	8921.52	695.09
27	Bus/Fin	34	5	Female	8976.72	537.01
28	IT/Eng	34	2	Female	7955.15	279.04
29	Bus/Fin	34	5	Female	10313.68	686.41
30	IT/Eng	35	2	Female	7380.6	500.96
31	IT/Eng	35	1	Female	9686.88	271.49
32	IT/Eng	35	3	Female	6945.16	419.45
33	IT/Eng	35	3	Female	10130	710.38
34	IT/Eng	35	1	Female	9454.8	519.65
35	IT/Eng	35	5	Female	6875.78	454.07
36	IT/Eng	36	1	Female	7940.91	749.45
37	IT/Eng	37	5	Female	9186.56	723.74
38	IT/Eng	37	5	Female	10041.2	568.12
39	IT/Eng	37	4	Female	9498.4	501.65
40	Others	39	1	Female	9218.24	518.06
41	Bus/Fin	39	3	Female	8553.44	577.6
42	IT/Eng	39	6	Female	7666.1	460
43	Others	39	1	Female	11288.16	471.31
44	Others	39	4	Female	7056.47	555.35
45	Bus/Fin	39	5	Female	10167.28	398.03
46	Bus/Fin	40	3	Female	8024.72	547.03
47	IT/Eng	40	6	Female	7081.9	536.28
48	IT/Eng	40	4	Female	9330.32	500.5
49	Bus/Fin	41	1	Female	8336.96	444.38
50	IT/Eng	41	4	Female	7603.2	419.53
51	Others	42	4	Female	5028.62	621.32
52	IT/Eng	42	3	Female	7378.14	466.73
53	IT/Eng	43	6	Female	8563.44	508.67
54	IT/Eng	43	2	Female	10868.96	296.47
55	IT/Eng	44	5	Female	8351.12	557.35
56	Others	44	1	Female	9864.56	310.47
57	Others	44	3	Female	7546.02	460.37
58	Others	44	6	Female	8315.12	130.03
59	Others	44	3	Female	7977.22	536.22
60	IT/Eng	44	4	Female	7640.8	606.88
61	IT/Eng	44	6	Female	9520	437.9
62	IT/Eng	45	6	Female	10125.68	633.5
63	Others	46	1	Female	7417.85	565.32
64	IT/Eng	46	4	Female	7405.18	345.18
65	IT/Eng	48	3	Female	6864.86	335.57
66	IT/Eng	48	4	Female	9822.88	523.66
67	Others	48	4	Female	7775.92	504.83
68	IT/Eng	49	6	Female	9200	305.35
69	IT/Eng	49	5	Female	9683.44	546.46
70	Others	50	1	Female	8175.84	545.24
71	Bus/Fin	50	5	Female	7151.78	623.88
72	Others	52	4	Female	12000	423.39
73	IT/Eng	52	6	Female	7570.6	397.94
74	IT/Eng	52	5	Female	8558.88	479.93
75	Bus/Fin	52	2	Female	7681.66	799.55
76	Bus/Fin	52	3	Female	7229.6	273.78
77	Others	52	5	Female	11091.04	438.76
78	Others	53	2	Female	9180.48	434.59
79	IT/Eng	53	3	Female	7735.74	501.85
80	IT/Eng	53	6	Female	8987.84	743.84
81	IT/Eng	53	4	Female	7340.88	460.21
82	Others	55	4	Female	7628.06	577.98
83	IT/Eng	55	3	Female	7301.79	480.58
84	IT/Eng	55	5	Female	9389.68	469.83
85	IT/Eng	55	1	Female	6739.38	303.77
86	IT/Eng	55	5	Female	7432.54	423.54
87	IT/Eng	57	6	Female	7804.9	572.23
88	IT/Eng	57	3	Female	6604.22	617.24
89	IT/Eng	57	1	Female	8014.4	386.85
90	IT/Eng	58	6	Female	7980.16	429.63
91	IT/Eng	58	1	Female	7912.84	421.49
92	Others	59	1	Female	9509.76	704.26
93	IT/Eng	59	6	Female	7341.42	530.9
94	IT/Eng	59	1	Female	7141.75	485.78
95	IT/Eng	60	4	Female	6775.33	572.48
96	IT/Eng	60	2	Female	8119.36	706.38
97	IT/Eng	60	2	Female	6147.52	384.43
98	Others	60	3	Female	9280.56	440.37
99	IT/Eng	60	1	Female	9950.08	572.09
100	Bus/Fin	60	5	Female	10594.08	439.01

No.	Job sector	Age	No. of visits per month	Gender	Household income (in $)	Amount spent per month (in $)
101	IT/Eng	61	1	Female	8937.2	604.36
102	IT/Eng	61	2	Female	6324.46	701.94
103	IT/Eng	61	1	Female	9111.84	406.34
104	IT/Eng	61	6	Female	6754.44	494.46
105	IT/Eng	61	3	Female	10053.84	493.23
106	Bus/Fin	61	1	Female	8960	583.01
107	IT/Eng	62	5	Female	8245.12	398.3
108	IT/Eng	62	5	Female	8577.28	595.97
109	IT/Eng	62	6	Female	8079.12	718.32
110	IT/Eng	62	6	Female	8575.68	448.86
111	Bus/Fin	20	3	Male	8078.08	615.07
112	Bus/Fin	22	4	Male	9205.28	718.42
113	Bus/Fin	23	6	Male	7703.13	557.55
114	Others	23	1	Male	6948.54	328.81
115	Bus/Fin	24	6	Male	10140.16	400.01
116	Bus/Fin	24	2	Male	8066.56	380.55
117	Bus/Fin	24	5	Male	7931.98	633.94
118	Bus/Fin	25	3	Male	7876.96	623.06
119	Bus/Fin	25	6	Male	8068.16	422.38
120	IT/Eng	27	5	Male	7538.61	458
121	Bus/Fin	27	5	Male	9517.76	714.08
122	IT/Eng	27	2	Male	8850.88	578.7
123	Bus/Fin	27	2	Male	6702.81	759.77
124	Bus/Fin	28	2	Male	9738.88	547.47
125	Bus/Fin	28	6	Male	8095.68	435.69
126	Bus/Fin	28	3	Male	7918.79	150.35
127	Bus/Fin	29	3	Male	6955.31	516.49
128	Bus/Fin	29	6	Male	8203.28	357.14
129	Bus/Fin	29	1	Male	8062.32	697.92
130	IT/Eng	29	6	Male	8370	498.49
131	IT/Eng	30	1	Male	5924.37	377.35
132	Bus/Fin	30	2	Male	7834.53	626.42
133	Others	31	5	Male	6065.54	708.58
134	IT/Eng	32	2	Male	8617.44	629.97
135	Bus/Fin	32	6	Male	11016.16	617.78
136	Others	33	4	Male	5882.7	565.15
137	IT/Eng	33	2	Male	7315.18	454.91
138	Bus/Fin	33	4	Male	9264.24	470.76
139	IT/Eng	33	4	Male	9554.32	757.38
140	Bus/Fin	34	2	Male	6645.55	106.67
141	Bus/Fin	34	6	Male	7828.06	714.31
142	IT/Eng	35	1	Male	7455	367.63
143	Bus/Fin	35	3	Male	8718.88	461.5
144	IT/Eng	35	4	Male	8168.56	358.09
145	IT/Eng	35	1	Male	8322.48	771.98
146	Bus/Fin	36	3	Male	7439.84	365.68
147	IT/Eng	36	4	Male	7603.98	505.87
148	IT/Eng	36	5	Male	6453.19	439.28
149	Bus/Fin	36	6	Male	10400	513.48
150	Bus/Fin	37	5	Male	7334.46	659.28
151	IT/Eng	37	1	Male	8349.68	303.19
152	IT/Eng	38	6	Male	7071.22	457.92
153	IT/Eng	38	1	Male	8602.64	591.45
154	Others	39	5	Male	7326.22	258.44
155	Others	40	5	Male	8346.56	606
156	IT/Eng	40	2	Male	8021.2	504.37
157	IT/Eng	40	5	Male	8092.08	689.16
158	IT/Eng	40	4	Male	9600	514.32
159	Others	41	2	Male	7272.61	270.5
160	Others	41	3	Male	7477.06	412.38
161	Bus/Fin	42	5	Male	8290.88	650.01
162	IT/Eng	43	5	Male	8746.64	544.08
163	Bus/Fin	43	2	Male	8697.12	508.22
164	Others	43	4	Male	10144	475.03
165	Others	44	5	Male	7223	766.51
166	IT/Eng	44	4	Male	10484.48	398.39
167	Others	44	1	Male	7433.19	419.39
168	Others	44	1	Male	8571.12	683.45
169	IT/Eng	44	6	Male	9293.76	573.89
170	IT/Eng	45	5	Male	9772	526
171	IT/Eng	45	1	Male	7750.62	564.51
172	Others	45	3	Male	8265.44	385.24
173	Bus/Fin	45	4	Male	6181.74	763.24
174	IT/Eng	45	4	Male	6980.31	562.92
175	IT/Eng	46	2	Male	12000	773.94
176	IT/Eng	48	6	Male	8324.96	431.06
177	Others	48	6	Male	8492.32	214.46
178	IT/Eng	48	3	Male	10422.72	655.75
179	IT/Eng	49	3	Male	8199.76	326.25
180	Others	50	6	Male	9836.88	539.39
181	IT/Eng	50	6	Male	7995.7	525.93
182	Others	51	4	Male	6206.42	665.62
183	Others	51	4	Male	6791.86	609.67
184	Bus/Fin	51	2	Male	8369.52	561.95
185	IT/Eng	51	6	Male	6914.22	400.74
186	IT/Eng	51	2	Male	6426.92	331.43
187	IT/Eng	52	2	Male	8118.24	513.57
188	Others	52	1	Male	8008.48	660.83
189	Bus/Fin	54	3	Male	8881.44	355.79
190	IT/Eng	54	4	Male	9967.12	314.15
191	IT/Eng	55	4	Male	10329.04	397.82
192	IT/Eng	57	2	Male	8080.88	483.82
193	IT/Eng	57	2	Male	7289.73	484.15
194	Others	59	5	Male	8378.32	52.98
195	IT/Eng	59	1	Male	8499.84	616.73
196	Bus/Fin	60	1	Male	6616.04	580.12
197	IT/Eng	60	2	Male	7268.83	441.53
198	IT/Eng	60	6	Male	7450.53	521.33
199	IT/Eng	61	5	Male	6089.39	526.55
200	IT/Eng	61	5	Male	8091.92	460.2

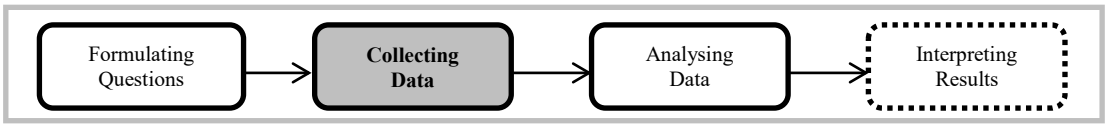


3.2 Common Statistical Terms

It is not uncommon to handle large amount of data in statistics. But with such large amount of data, what statistical techniques are there to churn these _____ into _____?

Let us first define some terminologies in statistics, shown as follows, before we look at types of data in statistics:

Terminology	Definition	Example from the Case Study
Variable	A quantity that can be measured and may take on different values within a problem.	
Data	Observations or responses collected for the selected variable. (A single observation is called *datum*.)	
Population	The complete set of items which we are studying. This is usually too large for the collection of data.	
Population Size	The number of items in the population.	
Sample	A subset of items selected from the population. When the population is too large, a representative sample is usually selected instead.	
Sample Size	The number of items in the sample.	



3.3 Types of Data

There are mainly two types of data – **qualitative** and **quantitative**.

_____ data are _____ values that are descriptive in nature. It is often used interchangeably with the term **categorical** data.
_____ data take on values measured on a _____ scale.

Fill in the heading (*Qualitative*/*Quantitative*) for the following data:

number of children height weight number of heads in coin tosses	colour shoe size blood type exam grade

Qualitative data can be further classified by **nominal** or **ordinal** scale.
Nominal scale data are identified by names or labels only, whereas ordinal scale data can be ordered or ranked.

Fill in the heading (*Nominal*/*Ordinal*) for the following qualitative data:

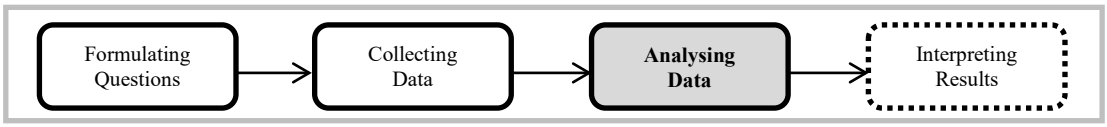
shoe size exam grade	colour blood type

Quantitative data can be further classified into **discrete** or **continuous** data.
Discrete data can only take on certain values, whereas continuous data can take on any value within a range. We say that discrete data is counted, whereas continuous data is measured.

Fill in the heading (*Discrete*/*Continuous*) for the following quantitative data:

number of children number of heads in coin tosses	height weight

Sketch a map of the types of data here:



4. Analysing Data

4.1 Summarizing Data

How do we describe a set of data? We can group them and present their pattern or distribution in a tabular or graphical form. We can also describe data by using a few well-chosen numbers that summarise meaningfully the entire data set. Hence, we can summarize the data, in two ways – by **graphical summary** and by **numerical summary**.

Nowadays, your calculators are equipped with statistical functions which enable almost effortless computations of the numerical summaries. Furthermore, software packages are able to produce sophisticated graphs easily. In this course, you will learn how to produce the numerical summaries and generate graphs using the statistical software **Minitab**. As such, the focus will be to learn how to interpret the summaries, rather than the "formulae" behind the summaries.

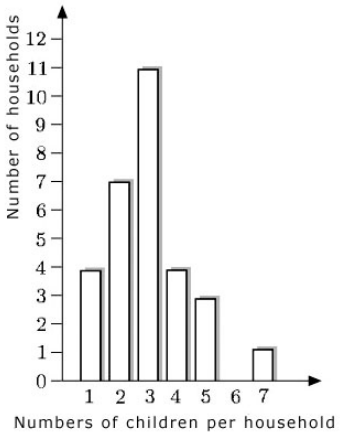
4.2 Graphical Summaries

An effective way to present a set of data to a team of decision makers is to use diagrams or graphs. Pattern exhibited by a variable and comparisons between variables become visual.

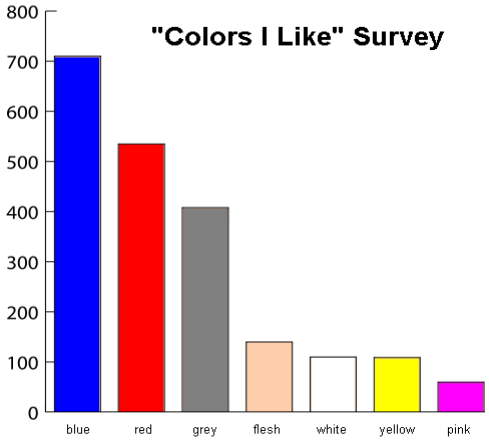
In this course, we will cover the more commonly used graphs – bar graph, pie chart, histogram and box plot.

- **Bar graph**

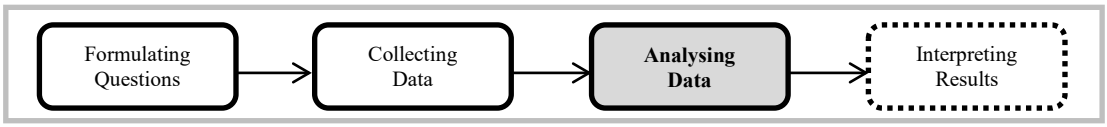
  Typically used to represent _____ or _____ data. It gives a visual overview of differences in _____ (or percentage) between categories.



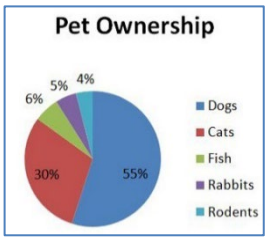
An example bar graph of a quantitative variable (i.e. number of children per household)



An example bar graph of a qualitative variable (i.e. favourite colour)



- **Pie chart**

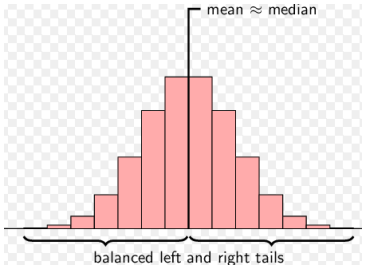
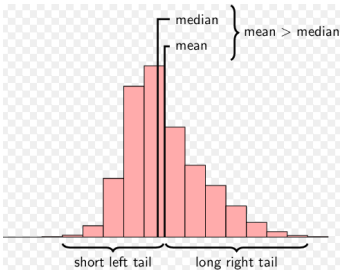
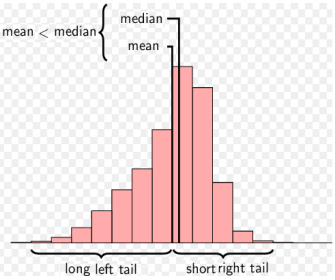


Typically used to represent _____ data.
It gives a visual overview of _____ belonging to each category.

- **Histogram**

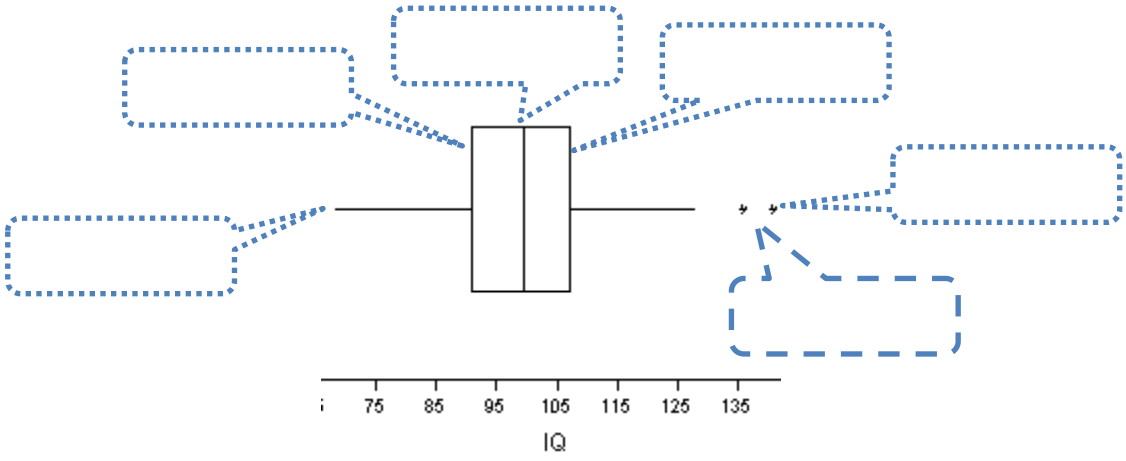
A histogram displays frequencies of _____ data that have been sorted into intervals. These give visual overview of the _____ of distribution of the data values. Specifically, _____ is a measure of symmetry of the data distribution, or rather, asymmetry.

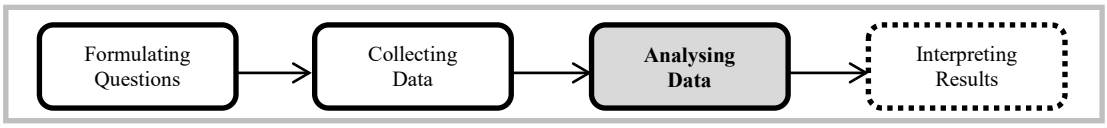
Histogram is similar to a bar chart in that they both use bars, either horizontal or vertical, to represent the number of data points in each category or interval. However, a histogram has no spaces between bars.

Symmetrical distribution	**Positively-skewed** distribution	**Negatively-skewed** distribution
		
In such cases, we can see (approximately) mirror images about the _____. Note that mean ≈ median.	The tail extends to the _____. Note that mean > median.	The tail extends to the _____. Note that mean < median.

- **Box plot**

Also known as **box and whiskers plot**, is another way to display _____ data. It is especially effective for <u>comparing</u> multiple groups of data sets.
We will need to generate a _____ in order to construct boxplot.





Minimum value
The _____ data value in the data set.

First quartile, Q1
___ percentile; can be found by finding median of the lower half of the data set.

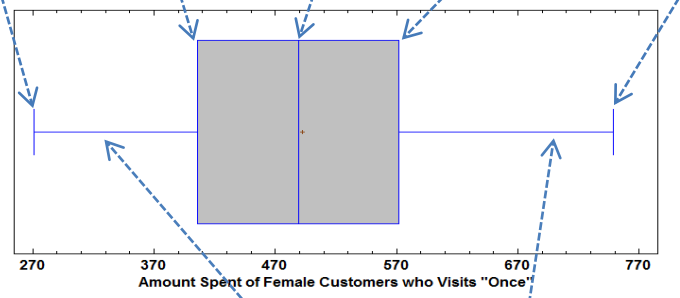
Third quartile, Q3
___ percentile; can be found by finding median of the upper half of the data set.

Sort the data set in _____ order.

Median, Q2
___ percentile; it will divide the data set into two halves.

Maximum value
The _____ data value in the data set.

Once	271	304	310	387	406	421	444	471	486	489	518	520	545	565	572	583	604	704	749



It is essential that the box plot is constructed to scale.

The box plot can be constructed on a horizontal or vertical scale.

Any **outliers** should be marked with an _____ or an open dot.
(You will be taught how to identify outliers next)

If one whisker is longer than the other, the distribution of the data is likely skewed in the direction of the _____ whisker.
(In this case, the distribution is skewed to the _____)

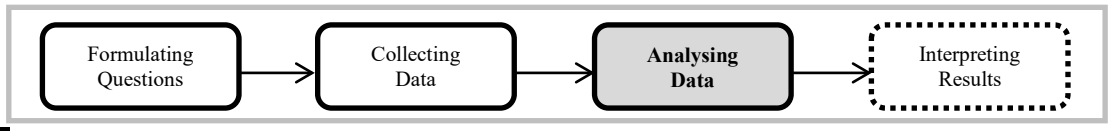
The box plot shows much of the structure of the data at a quick glance:

- the centre
- two measures of spread (interquartile range and range)
- skewness
- existence of outliers

To identify outliers, we compute the values of the **fences**:

- **Lower fence** can be calculated by the formula: Q1 – 1.5×IQR
  In the "Once" data set: lower fence =
- **Upper fence** can be calculated by the formula: Q3 + 1.5×IQR
  In the "Once" data set: upper fence =

Any extreme data values that fall outside the fences are considered to be **outliers**.
Note that fences are not indicated in the box plot.
In the "Once" data set: since all data values fall within the fences, there is no outlier.



4.3 Numerical Summaries

We can describe data by using a few well-chosen numbers that summarise meaningfully the entire data set.

Typically, it is useful to know where the centre or middle of the data set is, referred to as **measures of centre**. It is also known as measures of central tendencies or measures of central location. This is a single value that best represents the concentration of data, and suggests the "average" value of a distribution.

However, measures of centre alone provide only a partial description of a data set. We need a measure to indicate the spread or variation of quantitative data values. These measures are called **measures of dispersion**. In fact, these measures are of essential importance in statistics which is, mainly, the study of variability.

The various measures of centre and dispersion are listed here:

Measures of Centre	Measures of Dispersion
Mode **Median** **Mean**	**Range & Interquartile Range** **Standard Deviation & Variance**

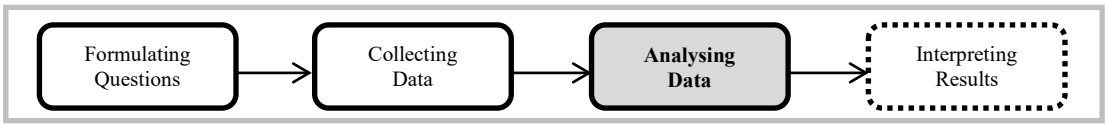
The selection of the appropriate measures of centre is described in the table below:

Measures	Mode	Median	Mean
Method	The most-likely occurring data value.	The centre or middle data value.	The numerical average of the data values.
Application	Most useful in, but not limited to, _____ data.	Good for _____ data that has outliers and/or is skewed.	Good for _____ data that are quite symmetrical and has no outlier.

Furthermore, comparing the values of these measures of centre (usually mean and median suffice) gives a quick sense of the distribution of the data in terms of **skewness**.

Distribution	Negatively-skewed	Symmetrical	Positively-skewed
Comparison		mean = median = mode	

Further elaboration on each measure of centre follows…



- **Mode**
  - The mode of a data set is the data value that occurs with the _____ frequency.
  - If all data values have same frequencies, then the data set has no mode.
  - If two data values occur with the same greatest frequency, then both the data values are considered modes. Such data with two modes are known as bimodal.

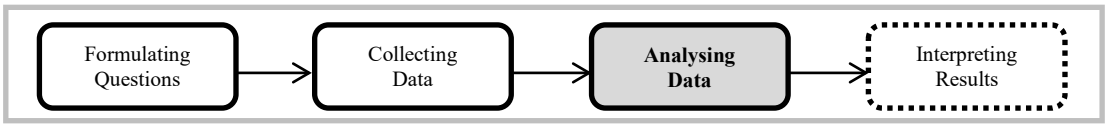
  Examples:
  5, 8, 13, 15, 17 3, 5, 7, 13, 3, 7, 9, 3 1, 1, 2, 2, 2, 2, 3, 4, 5, 5, 5, 5, 6, 7, 9

- **Median**
  - The median of a data set is the value that lies in the _____ when the data set is ordered. It is also known as the **second quartile** (**Q2**) or the **50th percentile**.
  - If the data set has an even number of observations, then the median is the midpoint of the two middle data observations.
    If the data set has an odd number of observations, then the median is the middle data observation.
  - The median is not influenced by extreme data values.
  - In addition, since half of the data values fall below Q2 and the other half of the data values fall above Q2, the median of the lower half of the data values is known as **lower quartile** or **first quartile** (**Q1**) or **25th percentile**.
  - Similarly, the median of the upper half of the data values is known as **upper quartile** or **third quartile** (**Q3**) or **75th percentile**.

  Examples: 4, 7, 9, 11, 12, 20 5, 8, 10, 10, 15, 18, 99

- **Mean**
  - This is the most popular and arguably, most accurate measure of centre.
  - Its value is obtained by "levelling out" the entire data set, hence every data value is used.
  - As a result, mean can be heavily influenced by extreme data values.
  - Mean is meaningless as a measure for qualitative data.
  - The notation for _____ is $\boldsymbol{\mu}$ and for _____ is $\bar{\boldsymbol{x}}$.

  Example: 16, 17, 10, 13, 20, 18, 13, 14, 18

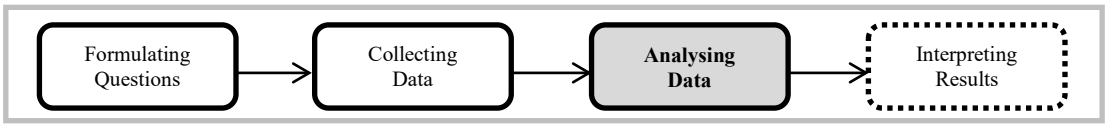


To measure spread or variation of quantitative data, further elaboration on each measure of dispersion follows…

- **Range & Inter-Quartile Range**
  - The **range** of a data set is simply the difference between the _____ and the _____ data values. Although it serves as a quick and easy measure of variability, it might not reflect the typical variability if either the largest or smallest (or both) data value is an extreme data value.
  - **Inter-quartile range** is the difference between the lower and upper quartiles, that is, IQR = _____. Since it measures variation of data values in the middle 50% of the data set, hence it is not affected by extreme data values.
  - Nevertheless, both range and inter-quartile range are based on only two data values in the whole data set. It does not reveal any information about the dispersion of the rest of the data values.

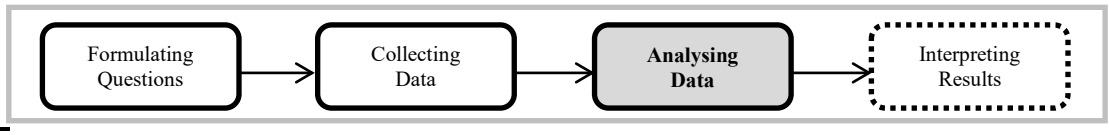
    Examples: 3, 4, 6, 7, 9 15, 15, 20, 25, 25, 30, 30, 30, 35, 75, 85

- **Standard Deviation & Variance**
  - The **standard deviation** is considered a more powerful measure of dispersion because it takes into account every data value in the data set, by summarising the amount by which each data value deviates from the mean.
  - Effectively, it indicates how tightly the data values in the data set are "bunched" around the mean value.
  - A _____ standard deviation implies that the data values are tightly bunched together, whereas a _____ standard deviation implies that the data values are spread apart.
  - The notation for _____ is $\boldsymbol{\sigma}$ and for _____ is $\boldsymbol{s}$.
  - **Variance** is mathematically the square of standard deviation. It represents the average squared deviation from the mean of the data.
  - The notation for **population variance** is $\boldsymbol{\sigma^2}$ and for **sample variance** is $\boldsymbol{s^2}$.



The selection of the appropriate measures of dispersion, paired with the corresponding measure of centre, is described in the table below:

Measures of Dispersion	**Inter-Quartile Range & Range**	**Standard Deviation & Variance**
Application	Range is a quick and easy measure but sensitive to outliers; whereas IQR is not sensitive to outliers. Both are good for _____ data.	Good for data that are quite _____. SD is more commonly used than its squared counterpart, variance.
Corresponding Measure of Centre	Median	Mean



4.4 Analysing Relationships

Let's use the scenario of investigating the **relationship** between motorboat propellers in Florida waterways and manatee fatalities from 1977 to 2011.

4.4.1 HOW CAN WE **VISUALIZE RELATIONSHIPS**?

The number of deaths and the number of powerboat registrations are both _____ variables. That means they can be measured numerically, and we can plot their values.

Instead of looking at a single variable, we can create a **scatter plot** to consider the relationship between these two variables.

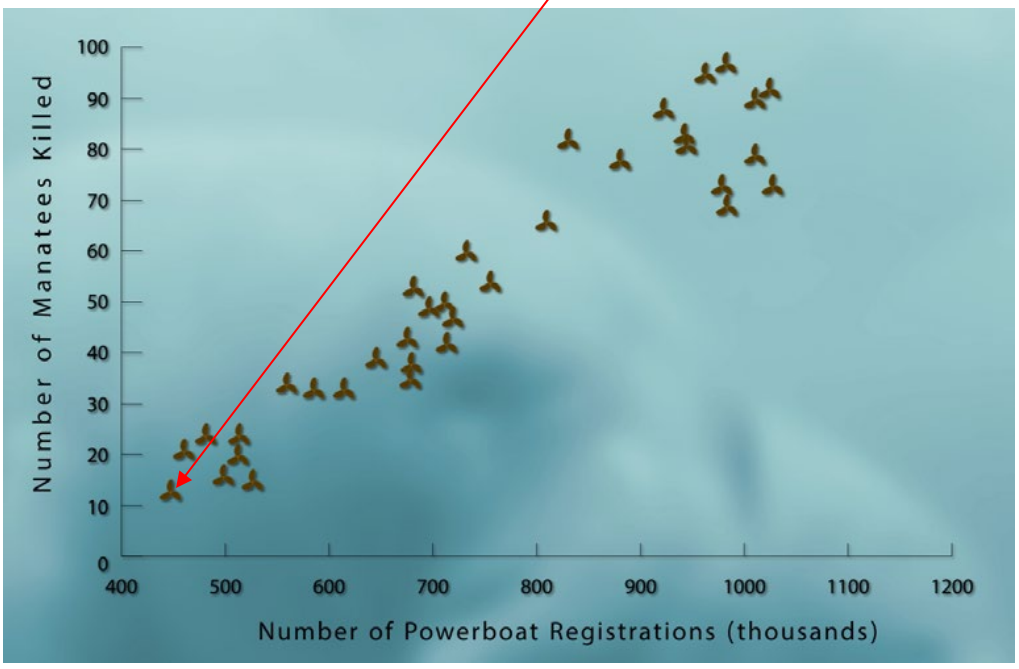
4.4.2 HOW DO WE **PRODUCE A SCATTER PLOT**?

To make a scatter plot, we first draw horizontal and vertical axes.

Since the number of powerboats in the water helps explain the number of manatees killed, thus the number of powerboat registrations is called the **explanatory variable**.
The explanatory variable always goes on the _____ axis.

We expect that the more boats that are in the water, the more manatees will be killed. That is, we assume that the number of manatees killed is a response to the number of boats in the water, thus we call the number of manatees killed the **response variable**.
The response variable always goes on the _____ axis.

Each point represents a datum. For example, the first point represents that (in 1977) the number of the registrations was 447,000 and the number of manatees killed by boats was 13.



4.4.3 WHAT DOES A **SCATTER PLOT SHOW**?

As the number of powerboat registrations increased, the number of manatees killed increased.
This is called a _____ association.
A _____ association would be when one variable increases while the other decreases.

The points roughly fall in a line. We call this pattern **linear**.
In fact, since the points do not deviate much from a line, we can say that the linear relationship is _____ between boats in the water and dead manatees.
If our data were all over the place with much deviations from the line, we would call the relationship _____.
However, not all relationships are linear; some show a curved pattern while some have no pattern at all.

When looking at scatter plot, we should look out for:

- Overall pattern – how strong it is and its direction
- Deviations from pattern
- Outliers

A scatter plot show the nature of a relationship between two variables, but it does not prove why the relationship exists. The changes in one variable do not necessarily *cause* the changes in the other; there could be other factors. **(Note: correlation does not imply causality.)**

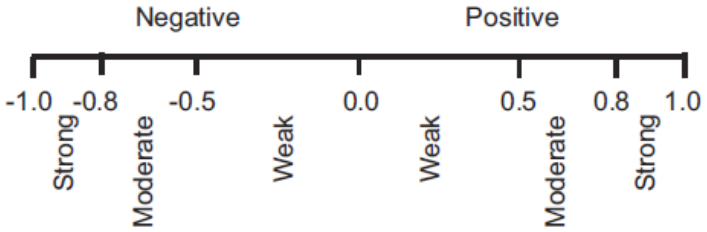
4.4.4 WHAT NUMERICAL SUMMARY CAN **MEASURE RELATIONSHIP**?

The sample **correlation coefficient**, denoted by ***r***, measures strength and direction of a linear relationship between two quantitative variables.

Basic properties of r :

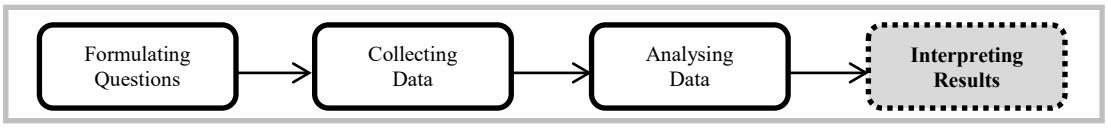
- The sign of r shows positive or negative association.
- The value of r always satisfies $-1 \le r \le 1$.
- The value of r remains the same when the two variables are interchanged or when the units of the variables are changed.

Guidelines on interpreting r :



Specifically:

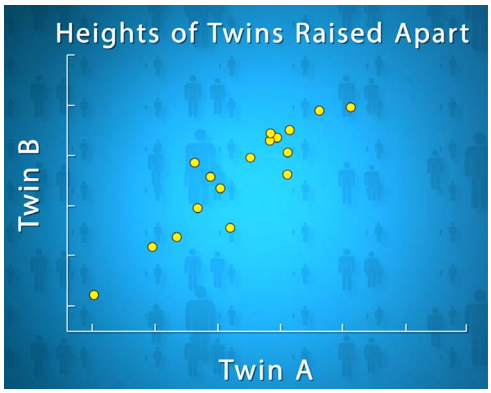
- Perfect positive correlation $\Rightarrow r =$ ___
- Perfect negative correlation $\Rightarrow r =$ ___
- No correlation $\Rightarrow r =$ ___



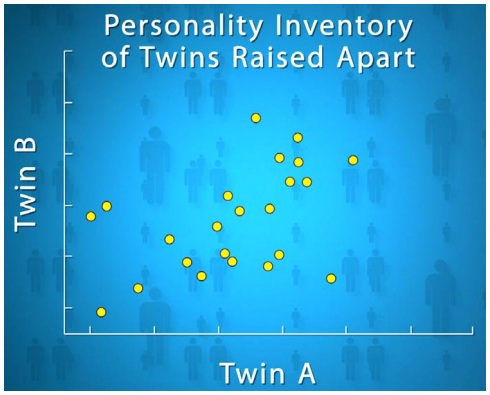
5. Interpreting Results

5.1 Interpreting Relationships

The following graphs and correlation values are produced from studying the physical and personality traits of identical twins who have been raised apart.



$r = 0.92$

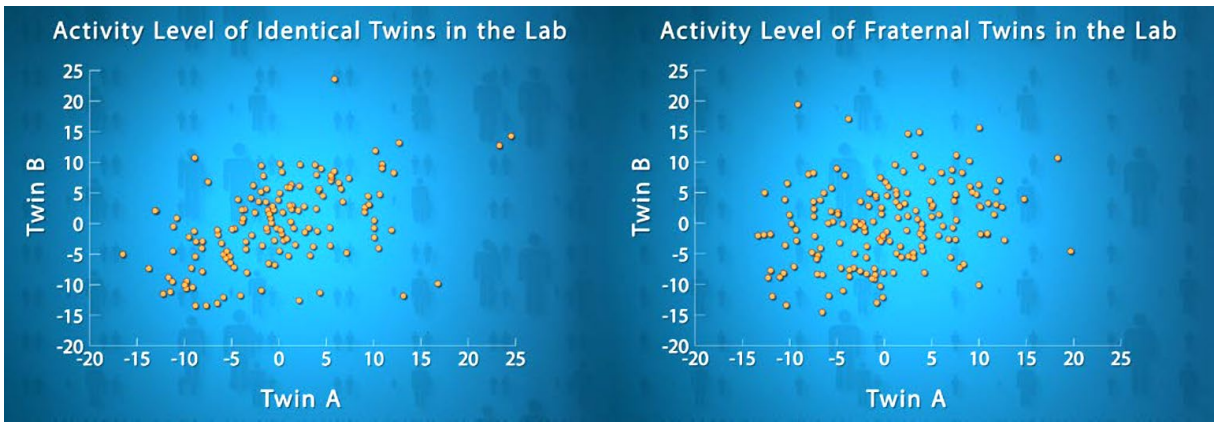


$r = 0.49$

We can observe the following:

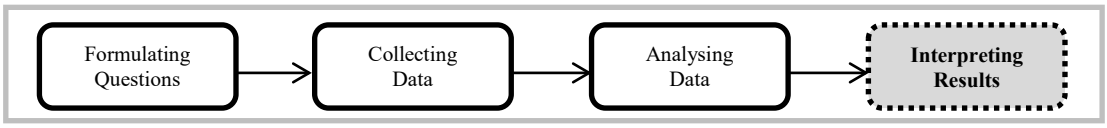
- From the plot on heights, the taller one twin is, the taller is the other. There is a positive association with strong pattern. Since $r = 0.92$, which is very close to 1, it indicates a strong, positive, linear association between heights of twins.
- From the plot on personality, though the relationship is not as clear as it was for height, the points do tend to increase together. Since $r = 0.49$, the relationship is not as strong as for height, but only moderate.

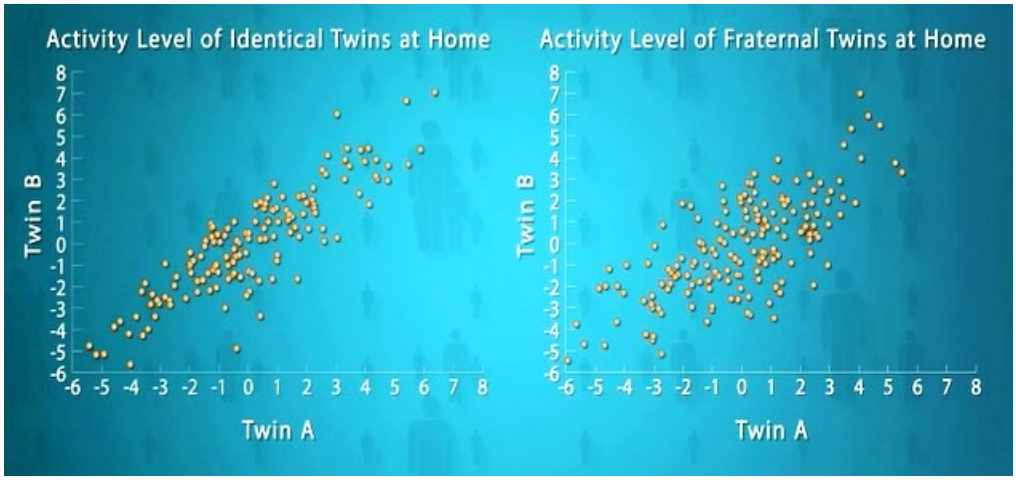
The following graphs and correlation values are produced from studying the activity level of twins in lab setting and at home; identical twins are on the left and fraternal twins on the right.



$r = 0.48$

$r = 0.26$





$r = 0.87$ $r = 0.7$

We can observe the following:

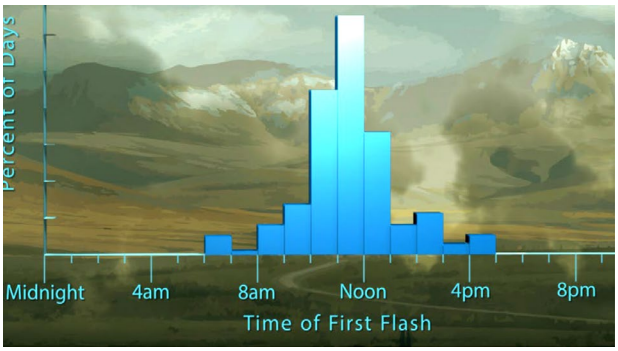
- In lab setting, there is moderate positive association between activity levels of identical twins, but weak positive association between activity levels of fraternal twins.
- Hence, in lab setting, the correlation between the activity levels of fraternal twins is much less than that between identical twins.
- In home setting, both plots show strong patterns.
- The correlation of activity levels in both identical and fraternal twins are much higher in the home setting (moderate to strong) than in the lab setting (weak to moderate).

5.2 Interpreting Graphical Summaries

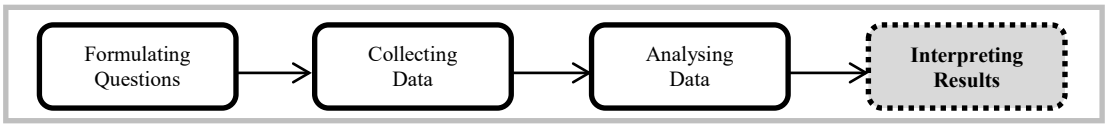
This histogram shows the time of first lightning strike collected over a particular year in a small area of Colorado, US.

We can observe the following from the graph:

- horizontal axis represents time of day
- vertical axis represents percentage of days
- each bar represents one hour
- roughly symmetrical about the tallest bar between 11am and 12 noon
- data is tightly clustered around the central bar, between 10am to 1pm
- no first strikes at night



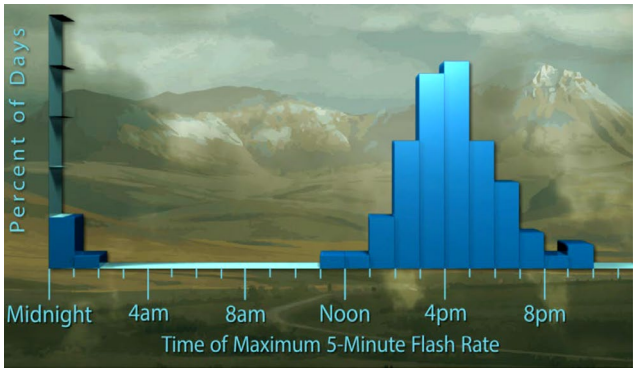
This histogram shows the time of day when the maximum number of lightning flashes (in 5 mins) were recorded in the same year and area as above.



We can observe the following from the graph:

- a peak shows that most flashes occur between 4pm and 5pm
- there are outliers where maximum flashes occur between 12am and 2am

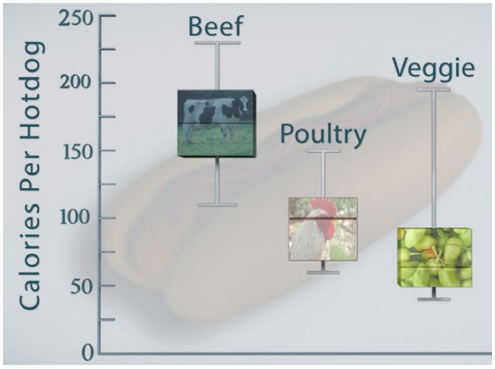
It is important when plotting a histogram to choose the best class size, that is, the width of intervals along the horizontal axis.



This box plot compares calories of beef, poultry and chicken hotdogs.

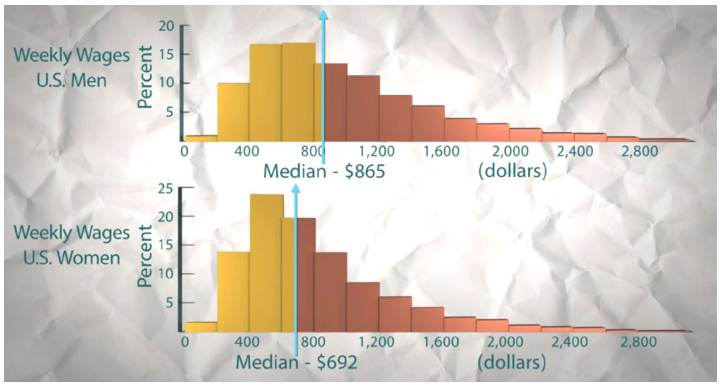
We can observe the following from the graph:

- The median of the poultry hotdogs lies below the minimum value for beef hotdogs, meaning the *typical* poultry hotdog has fewer calories than any beef brand.
- Overall, the veggie hotdogs have the lowest calories. But, the whiskers show that at least one veggie brand has more calories than ¾ of the beef hotdogs.



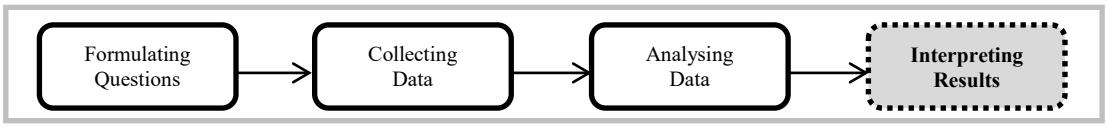
5.3 Interpreting Numerical Summaries

These histograms, marked with the respective medians, show the weekly wages of Americans in 2011, separated by gender.

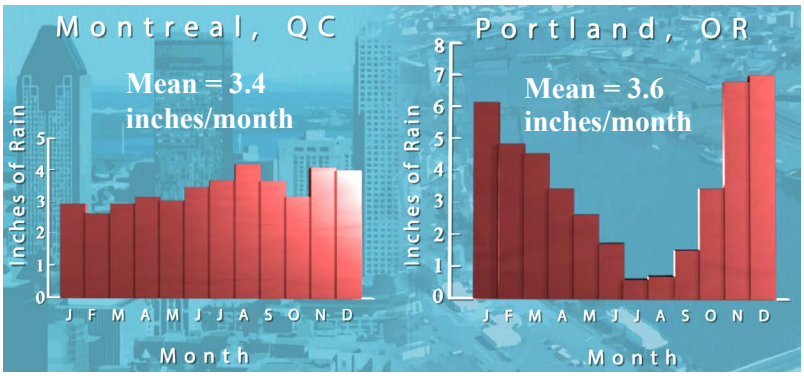


We can observe the following:

- Both histograms are skewed to the right, with most people making moderate salaries, while a few make much more.
- The median weekly salary for men in 2011 was $865. This means that half of all men made more than $865, and half earned less.
- The median wage for women was only $692, just 80% of what men make.



These graphs and statistics show the rain distribution of Montreal, Quebec and Portland, Oregon in a year.



We can observe the following:

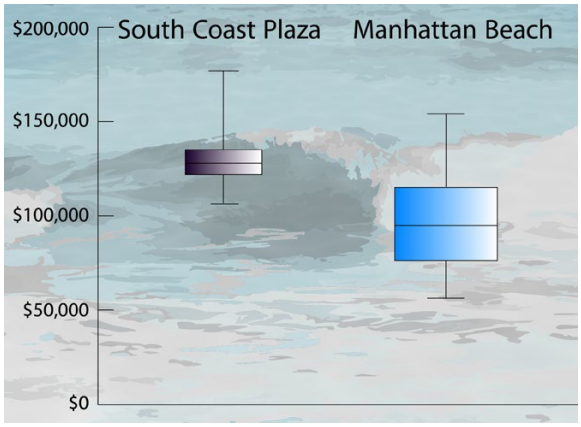
- The mean values show that average monthly rainfall for both cities are about the same, but they have very different climate
- From the graph, Montreal's rainfall is relatively consistent, measuring between 2 to 4 inches monthly.
- However, Portland's rainfall is much more varied, concentrated in the winter months, which can get almost 7 inches of rain, while summer months get less than 1 inch.

These box plots and statistics show the sales from two Wahoo's Fish Taco stores over four-week periods, one located at South Coast Plaza and the other located at Manhattan Beach.

<u>South Coast Plaza</u>

Mean = $130,675 per month

SD = $17,164



<u>Manhattan Beach</u>

Mean = $97,429 per month

SD = $31,075

We can observe the following:

- From the boxplots: The median sales of South Coast Plaza location is higher than that of Manhattan Beach location. But the interquartile range (represented by the widths of the boxes) for Manhattan Beach location is wider than South Coast Plaza location.
- South Coast Plaza location has higher mean sales than Manhattan Beach location.
- The SD values also show that the sales for Manhattan Beach location has greater variability than South Coast Plaza location.
-

Statistical Problem-Solving Process
Case Study: Prestige Mall

Student Name	Student Number	Class
(#1)		
(#2)		

Please refer to the case study of "Prestige Mall" on the chapter of Descriptive Statistics. Minitab will be used to analyzed the data.

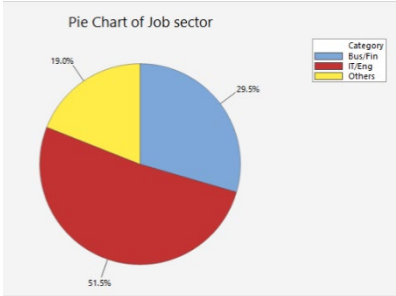
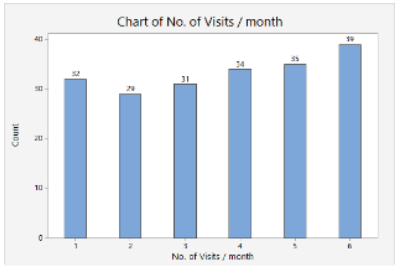
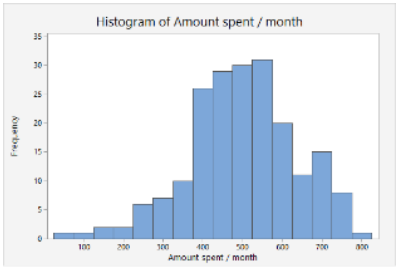
Q1: What is the aim of this case study?

Q2: What is the sample of this case study? And what is the targeted population?

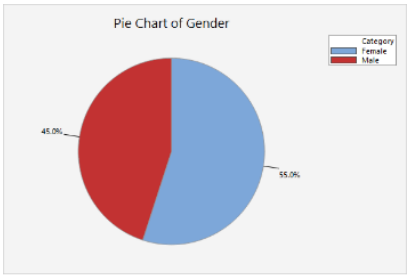
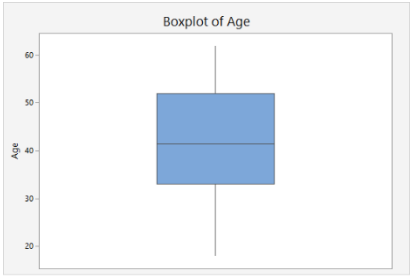
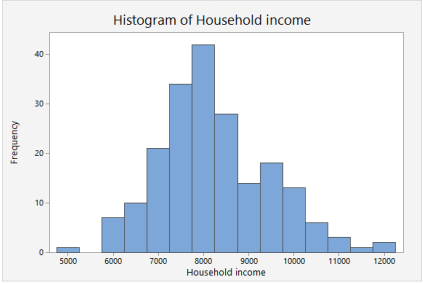
Q3: How were the data collected, as recorded in the data file named "Prestige Mall"?

Q4: What information (variables) does the data file named "Prestige Mall" hold?

Statistical Problem-Solving Process
Case Study: Prestige Mall

			Numerical and graphical summaries	Generalization to the target population
Q5	What is the proportion of customers in the IT/Eng and Bus/Fin sectors?	*Which variable data would you use to answer this question?* *What type of data is this?*	Bus/Fin: IT/Eng: Total proportion: 	*Describing the sample:* Most of the customers in Prestige Mall are professionals; about ____ are from BUS/FIN and IT/ENG. The mean number of visits is ____ and the median is __, which means that the customers frequent the mall about ___ times in the past month. So that is an average of about _____ a week. The mean and median amount spend last month is about _____. The amount spent is quite symmetrical, with SD of _____, clustering between _____ and _____. So, these customers surveyed spent about _____ to _____ in a week. *Generalizing to the population:* Generally, customers of Prestige Mall are mostly professionals who frequent the mall _____ a week, spending about _____ to _____ per visit.
Q6	How often do the customers visit Prestige Mall in the last month?	*Which variable data would you use to answer this question?* *What type of data is this?*	Mean: Median: SD: 	
Q7	How much did the customers spent last month at Prestige Mall?	*Which variable data would you use to answer this question?* *What type of data is this?*	Mean: Median: SD: 	

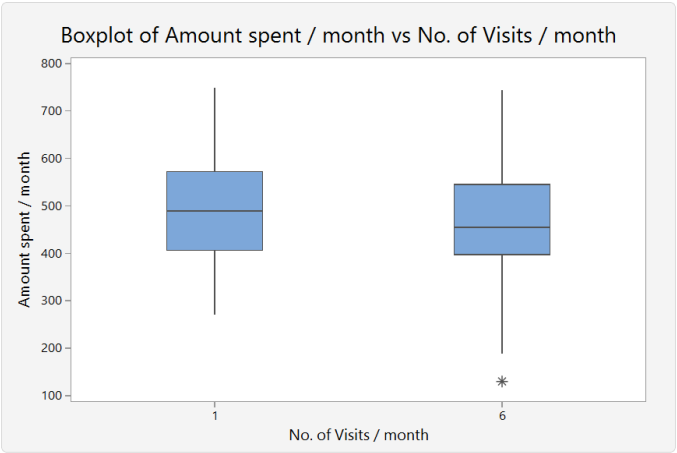
Statistical Problem-Solving Process
Case Study: Prestige Mall

			Numerical and graphical summaries	Generalization to the target population
Q8	What is the proportion of male and female customers of Prestige Mall?	*Which variable data would you use to answer this question?* **Gender** *What type of data is this?* **Qualitative (nominal)**	**Male: 45%** **Female: 55%** 	*Describing the sample:* There is a slightly higher proportion of female customers visiting Prestige Mall compared to male customers. The boxplot shows that the distribution of age of customers is quite symmetrical ranging from 18 to 62 years. The mean age is 42 years, with SD of about 12 years. Hence, customers who frequent Prestige Mall are more likely to be mature adults. The histogram for household income shows slight positive skewness. Many customers cluster around moderately low household income. The median household income is about $8k, with IQR of about $1.6k. *Generalizing to the population:* Generally, customers of Prestige Mall are mature adults, slightly more likely to be female, and could have moderate household income.
Q9	What is the age profile of the customers?	*Which variable data would you use to answer this question?* **Age** *What type of data is this?* **Quantitative (discrete)**	**Mean: 42.0 years** **Median: 41.5 years** **SD: 11.9 years** 	
Q10	What is the distribution of income of the customers?	*Which variable data would you use to answer this question?* **Household income** *What type of data is this?* **Quantitative (continuous)**	**Mean: $8231.29** **Median: $8067.36** **IQR: $1598.32** 	

Statistical Problem-Solving Process
Case Study: Prestige Mall

<table>
<tr><th></th><th></th><th></th><th>Numerical and graphical summaries</th><th>Generalization to the target population</th></tr>
<tr>
<td>Q11</td>
<td>Is there any preliminary evidence to claim that female customers who went to the mall 6 times spent more than female customers who went to Prestige Mall only once last month?</td>
<td>Which variable data would you use to answer this question?

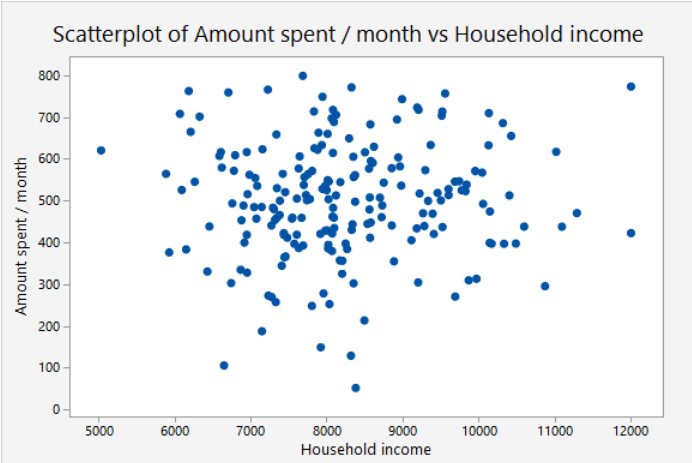
What type of data is this?

Which variable data is used for grouping?</td>
<td>
<table>
<tr><th>Amount spent/ month</th><th>Visit the mall once / month</th><th>Visit the mall 6 times / month</th></tr>
<tr><td>n</td><td>19</td><td></td></tr>
<tr><td>Mean</td><td>$ 492.24</td><td></td></tr>
<tr><td>SD</td><td>$ 127.24</td><td></td></tr>
<tr><td>Minimum</td><td>$ 271.49</td><td></td></tr>
<tr><td>Q1</td><td>$ 406.34</td><td></td></tr>
<tr><td>Median (Q2)</td><td>$ 489.26</td><td></td></tr>
<tr><td>Q3</td><td>$ 572.09</td><td></td></tr>
<tr><td>Maximum</td><td>$ 749.45</td><td></td></tr>
</table>


</td>
<td>Describing the sample:
The side-by-side boxplots showed that much of the boxes and whiskers _____. Also, there is an outlier value for female customers who visited the mall 6 times, indicating unusually low amount spent. The amount spent by female customers who visit once is typically about _____ than those who frequent 6 times (mean = $492 vs _____), and slightly more consistent (SD = $127 vs _____).

Generalizing to the population:
There is ____ visual evidence from the sample data to suggest that female customers who frequent the mall 6 times will spend more than those who visit once in a month.</td>
</tr>
</table>

Statistical Problem-Solving Process
Case Study: Prestige Mall

			Numerical and graphical summaries	Generalization to the target population
Q12	*(Ask a question about the relationship between 2 variables, then proceed to investigate.)* **Is there a relationship between…** _____ and _____ _____ by customers in Prestige Mall?	*Which variable data would you use to answer this question?* *What type of data are these?*	 Output from Minitab: *Pearson correlation of Household income and Amount spent / month =*	*Describing the sample:* The scatterplot _____ a linear relationship between the _____ _____ and _____ _____ by Prestige Mall customers. The correlation coefficient r = ____ is _____, indicating _____ _____ between the variables. *Generalizing to the population:* There is ____ evidence to suggest that _____ and _____ by customers of Prestige Mall are related.

Q13: Based on interpretations from Q5 to Q12, suggest a potential business to recommend to your boss, and thus a potential tenant (e.g Rolex?).
Note: You need not use all the interpretations.

TUTORIAL 1

1. To investigate the driving habits of Singaporeans, you would like to design a survey to collect data from a sample of 100 drivers.

   (a) Define the population and sample in this context.

   (b) Decide which of the following variables is relevant to your investigation and classify the type of data to be collected.

	Variable	**Relevant or not?**	**Type of data**
i	Age of driver		
ii	Height of driver		
iii	Weight of driver		
iv	Gender of driver		
v	Capacity of car (eg. 1600 cc)		
vi	Number of trips made per day		
vii	Distance covered per day		
viii	Amount of money spent on petrol per month		
ix	Colour of car		
x	Make (model) of car		
xi	Purchase price of car		

   (c) Select one of the relevant variables as indicated in part (b) and justify why this variable is relevant in your investigation.

   (d) Which type of graphs is suitable to present the data of the following variables?

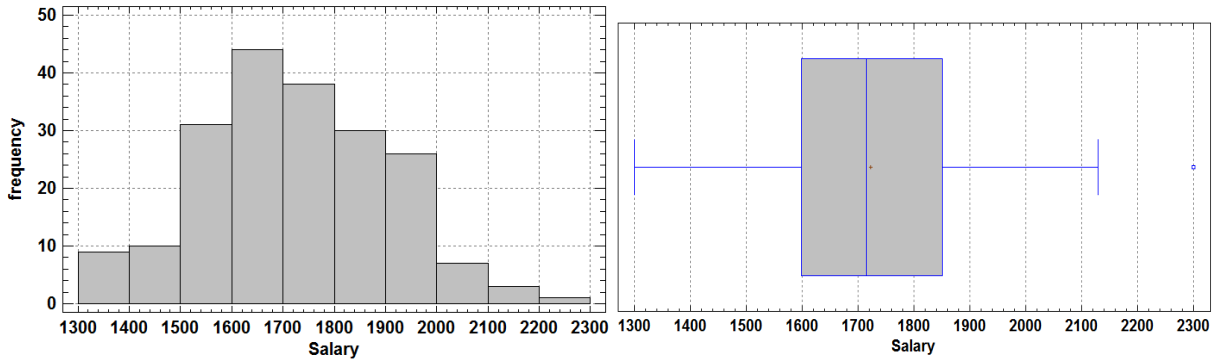
      I. Age of driver

      II. Gender of driver

      III. Number of trips made per day

   For each of the graphs selected to present the data in parts I to III, what information can be obtained from the graph?

2. Two hundred staff were randomly selected from a company and their salaries were presented using two charts, as shown in the following.



(a) What is the median salary of these 200 staff?

(b) Find the range and interquartile range of the salaries.

(c) What are the cut-off salaries for the bottom 25% and top 25% earners?

(d) Is there any outlier salary? What are the values of the fences?

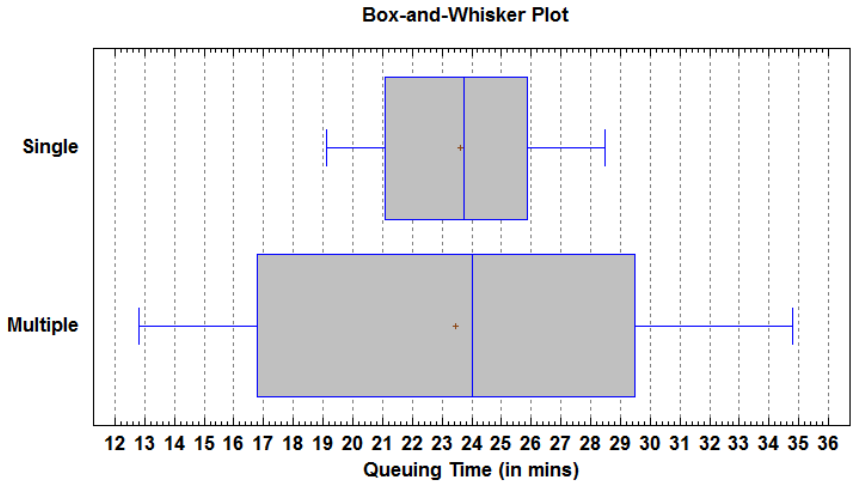
(e) How many staff earn between \$1800 and \$2000?

(f) Andrew earns \$1600. At which percentile is his salary?

(g) What is the shape of the distribution of salaries?

3. To serve customers better by cutting the queuing time at the counters, ABC Bank experimented with two types of queue system:
   - a single queue that feeds to all counters, or
   - multiple queues, one for each counter.

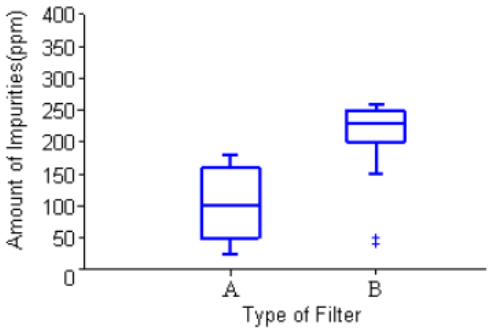
   The queuing times (in minutes) for 20 customers during the peak period before being served were recorded for each queue system. The results are displayed in the following box plots, where "+" inside the box represents the mean queuing time.



   Compare the two types of queue systems.

   Hints:
   - *Compare and comment on the measures of centre of both systems.*
   - *Compare and comment on the measures of dispersion of both systems, and discuss their pros and cons.*

4. Two new filtration systems A and B have been proposed for use in the water systems of a small city. The amount of impurities (in parts per million) remaining in the water after the water passes through each filter is recorded over a 30-day period. The average daily values for the two systems are plotted using a side-by-side box plot as follows:

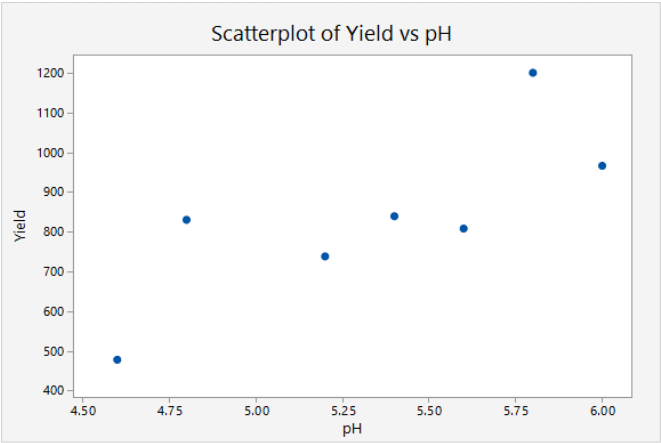


(a) For each filter, describe the shape of the distribution of the amount of impurities.
(b) Estimate the median, lower quartile and upper quartile for each filter.
(c) Which filter, A or B, produces less variability? Briefly explain.
(d) Which filter, A or B, appears to generally filter water more thoroughly?

5. A scientist planted alfalfa on several plots of land, identical except for the soil pH. The data collected and shown below give the yields (in kilograms per acre) for each plot. The scatterplot and correlation coefficient are also produced below.

pH	Yield
4.6	479
4.8	831
5.2	739
5.4	840
5.6	809
5.8	1201
6.0	967

$r = 0.78$



(a) Which is the explanatory variable and which is the response variable?
(b) Comment on the relationship between variables pH and yield.

6. You wish to compare the weight reducing program offered by two programmes, Program A and Program B. You have 60 participants and you randomly assigned thirty of them to each program. The data on the weight loss (in kg) of the participants two months after attending the programs were collected. Minitab gave the following summary:

Descriptive Statistics: Programme A, Programme B

Statistics

Variable	N	Mean	StDev
Programme A	30	4.0833	0.6086
Programme B	30	4.9633	0.5798

Which program is more effective in weight reducing? Explain.

ANSWERS

1. (a) Population: all Singaporean drivers
   Sample: the 100 Singaporean drivers surveyed
   (b) *<As long as you can justify, there is no correct or wrong answers to "relevance".>*
   (i) Quantitative (ii) Quantitative (iii) Quantitative
   (iv) Qualitative (v) Quantitative (vi) Quantitative
   (vii) Quantitative (viii) Quantitative (ix) Qualitative
   (x) Qualitative (xi) Quantitative
   (c) *<Sample answer>* For example, capacity of car: more powerful cars in the hands of amateur drivers may cause more reckless driving.
   (d) I. Histogram; to see the distribution of the age data
   II. Pie chart; to see the proportion of male and female drivers
   III. Bar chart; to see the differences between the number of trips recorded

2. (a) \$1720 (b) \$1000, \$250 (c) \$1600, \$1850
   (d) Yes, \$2300, LF = \$1225, UF = \$2225 (e) About 56 staff
   (f) About 25^{th} percentile (g) Slightly positively-skewed

3. The mean and median for both the system is approximately the same but the variation (as measured by the "box") of the multiple queue system is greater than that of single queue system. The minimum time for single system is higher than that of multiple queue system, but the maximum queue time for single system is lower than that of the multiple queue system. Although there is a possibility that a customer may be have a shorter queue time in a multiple queue system, but queue time for multiple queue system is not as consistent as single queue system.

4. (a) A is roughly symmetric; B is negatively-skewed with 2 outliers.
   (b) Filter A: Q1 ≈ 50 ppm , Q2 ≈ 100 ppm , Q3 ≈ 160 ppm
   Filter B: Q1 ≈ 200 ppm , Q2 ≈ 230 ppm , Q3 ≈ 250 ppm
   (c) Ignoring the outliers, B seems to produce less variability, as evident from the shorter width of the box, which represents an IQR of approximately 50 ppm.
   (d) Filter A

5. (a) Explanatory: pH; Response: Yield
   (b) Scatterplot shows positive association. Since $r = 0.78$, it indicates quite a strong positive linear relationship.

6. B, higher mean weight loss and more consistent weight loss (lower SD).

PRACTICAL 1 : Descriptive Statistics

Learning Objectives:

1. *Enter and import data into Minitab.*
2. *Generate numerical summaries using Minitab.*
3. *Generate graphical summaries using Minitab.*
4. *Generate correlation for bivariate data using Minitab.*

Task 1A

Input data into Minitab worksheet.

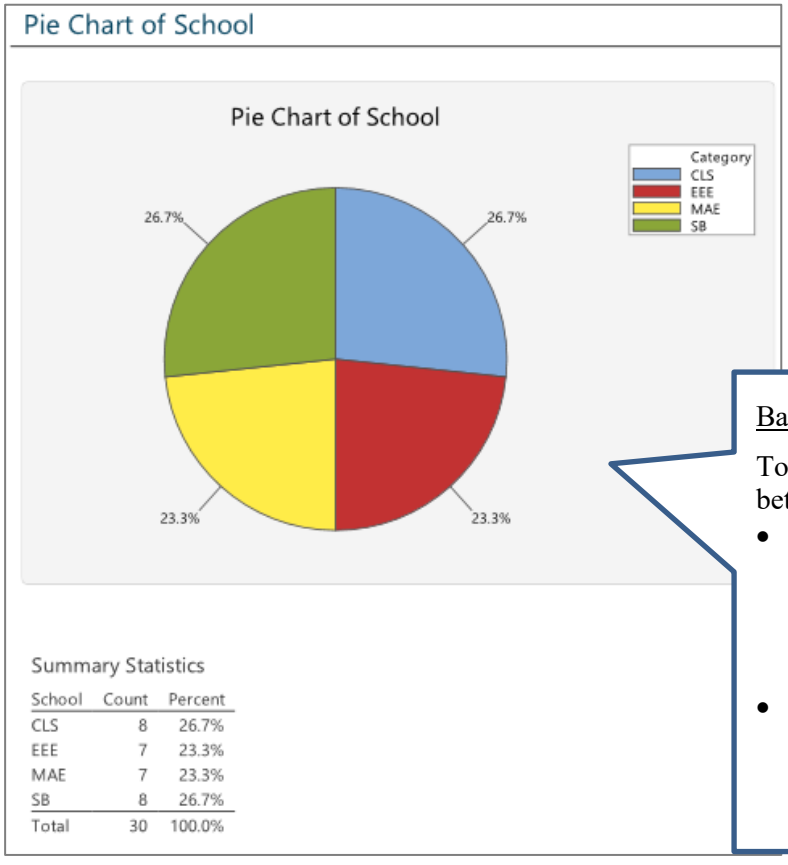
Copy data *School* and *GPA* from downloaded Excel file "*STAT_Prac1_Data.xlsx*".

Task 1B

Use the data set in "*Prac1*" Excel worksheet to construct various graphical summaries and provide basic interpretations. The graphs include:

- **pie charts**
- **bar graphs**
- **histograms**
- **boxplots**

(I) Construct a pie chart and a bar chart for the categorical variable *School*.
(II) Construct a histogram and a boxplot for the quantitative variable *GPA*.
(III) (Optional) Construct a histogram and boxplot for the quantitative variable *Starting Salary*.

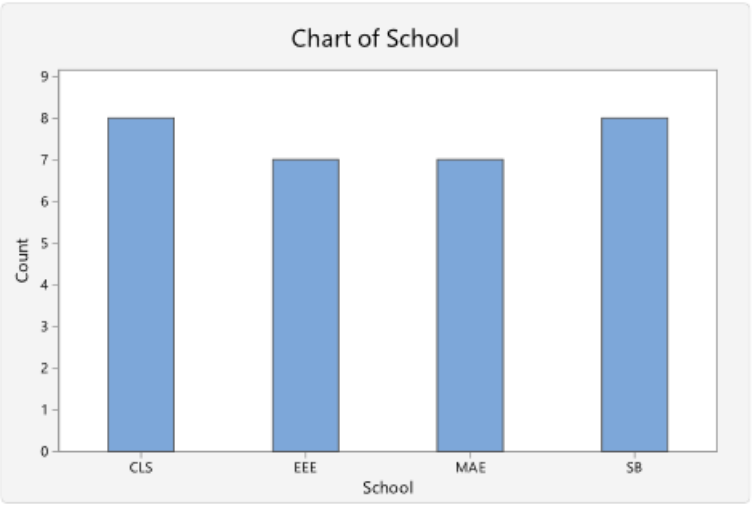


Basic Interpretations of Pie Charts

To interpret a pie chart, compare between groups.

- When you interpret single pie chart, look for differences in the size of the slices. The size of a slice shows the proportion of observations that are in that group.
- When you compare multiple pie charts, look for differences in the size of slices for the same categories in all the pie charts.

Chart of School



Summary Statistics

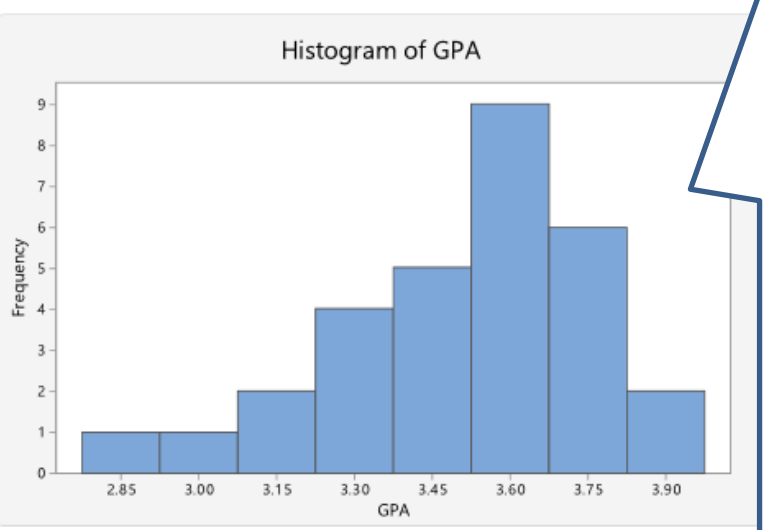
School	Count	Percent
CLS	8	26.7%
EEE	7	23.3%
MAE	7	23.3%
SB	8	26.7%
Total	30	100.0%

Basic Interpretations of Bar Charts

To interpret a bar chart, compare between groups.

- Look for differences in the heights of the bars.
- The bars show the value for the groups. Refer to the scale range of the *y*-axis to determine the actual differences.

Histogram of GPA



Summary Statistics

N	Mean	StDev	Minimum	Maximum
30	3.52200	0.24867	2.90000	3.92000

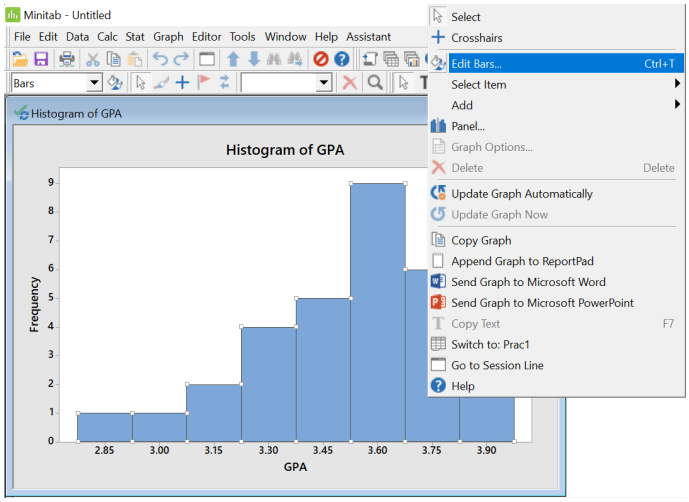
Basic Interpretations of Histogram

To interpret a Histogram, assess the key characteristics:

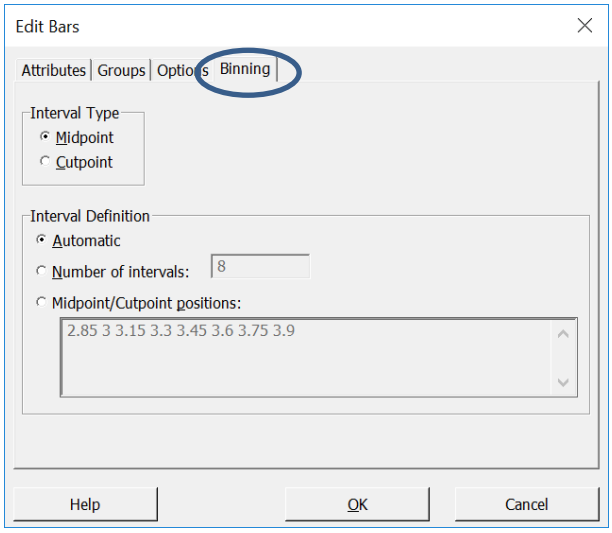
- Identify the peaks, which are the tallest clusters of bars. The peaks represent the most common values.
- Assess the spread of your sample to understand how much your data varies.
- Observe the skewness. When data are skewed, the majority of the data are located on the high or low side of the graph. Skewness indicates that the data may not be normally distributed.
- Outliers, which are data values that are far away from other data values, can strongly affect your results. Often, outliers are easiest to identify on a boxplot. On a histogram, isolated bars at the ends identify outliers.

CUSTOMIZING THE HISTOGRAM

Right-click on any bar. Select “Edit Bars”.



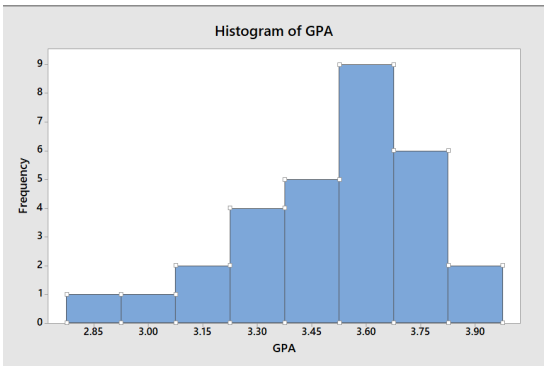
Click “Binning”.



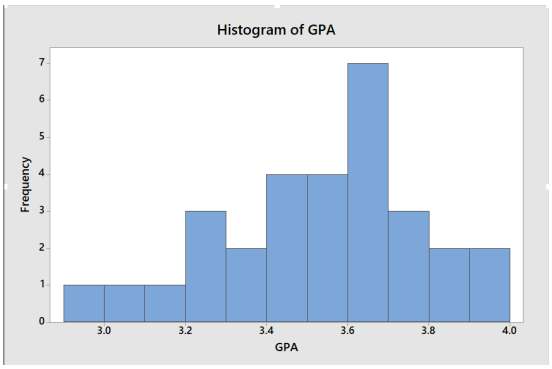
The following information describes some of the items on the **Binning** box:

Interval Type: Choose where to display the **tick labels**:

- Bins can be defined by either their midpoints (centre values) or their cut points (boundaries).
- The appearance of the graph changes if you change the bin definition method.



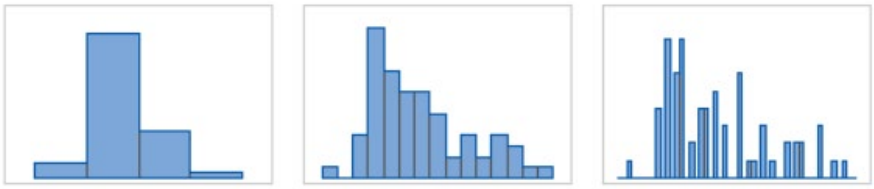
Midpoint



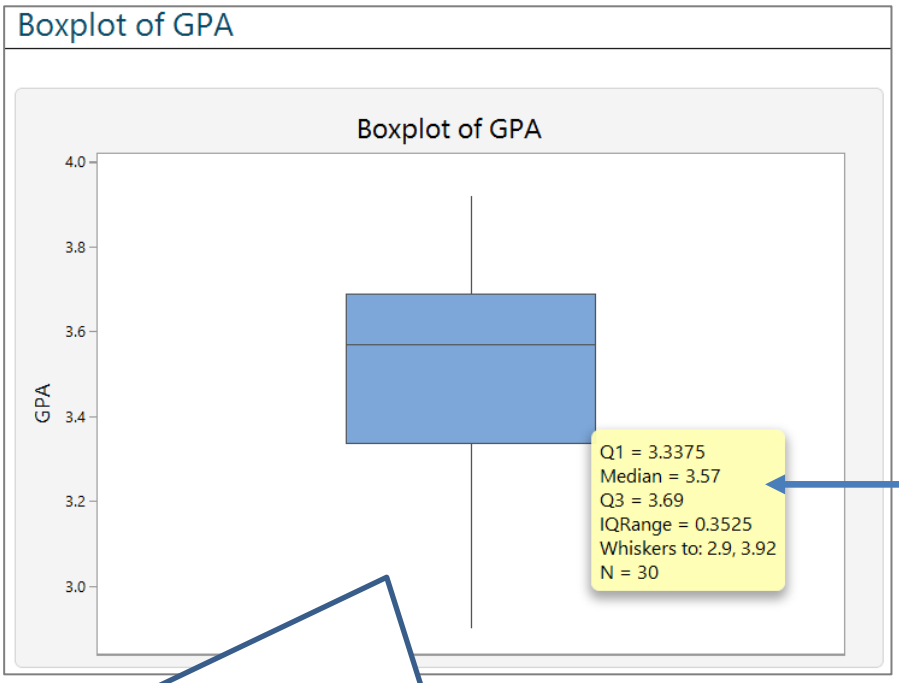
Cutpoint

Interval Definition: Choose the **number of intervals (bins)**:

- The number of bins affects the appearance of a graph. If there too few bins, the graph will be unrefined and will not represent the data well.
- If there are too many bins, many of the bins will be unoccupied and the graph may have too much detail. For example, these histograms represent the same data with different numbers of bins.



4 bins 15 bins 50 bins



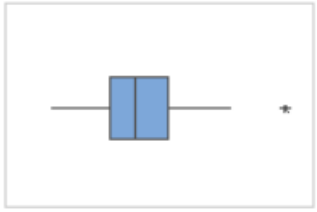
Hover the pointer over the boxplot to display a tooltip that shows numerical statistics.

Basic Interpretations of Boxplot

To interpret a boxplot:

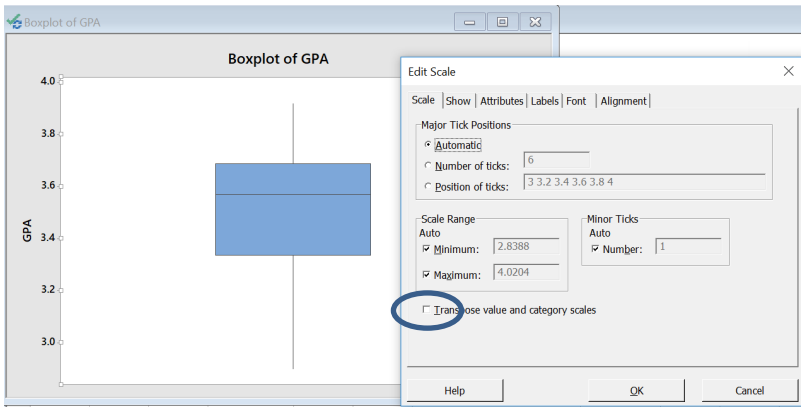
- Examine the following elements to learn more about the centre and spread of your sample data.
  - The median is represented by the line in the box. The median is a common measure of the centre of your data.
  - The interquartile range box represents the middle 50% of the data.
  - The whiskers extend from either side of the box. The whiskers represent the ranges for the bottom 25% and the top 25% of the data values, excluding outliers.
- Skewed data
  - When data are skewed, the majority of the data are located on the high or low side of the graph. Skewness indicates that the data may not be normally distributed (you will learn Normal distribution in Chapter 2).

- Outliers, which are data values that are far away from other data values, can strongly affect results. Often, outliers are easiest to identify on a boxplot. On a boxplot, outliers are identified by asterisks (*).

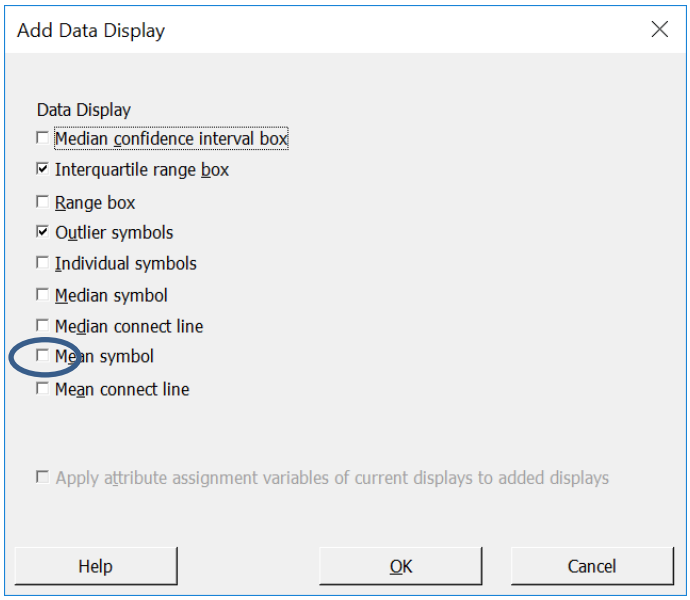


CUSTOMIZING THE BOXPLOT

To change a vertical boxplot to a horizontal boxplot, double-click on Horizontal or Vertical axis. Tick “Transpose value and category scales”.



To add a symbol for mean, choose “Editor” → “Add” → “Data Display”. Tick the “Mean symbol”



Task 1C

Use the dataset in "*Prac1*" Excel worksheet to compute numerical summaries of data.

Display summary statistics for the quantitative variable *GPA*.

Descriptive Statistics: GPA

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	Range	IQR
GPA	30	0	3.52200	0.04540	0.24867	0.06184	2.90000	3.33750	3.57000	3.69000	3.92000	1.02000	0.35250

Basic Interpretations of Numerical Summaries of Data

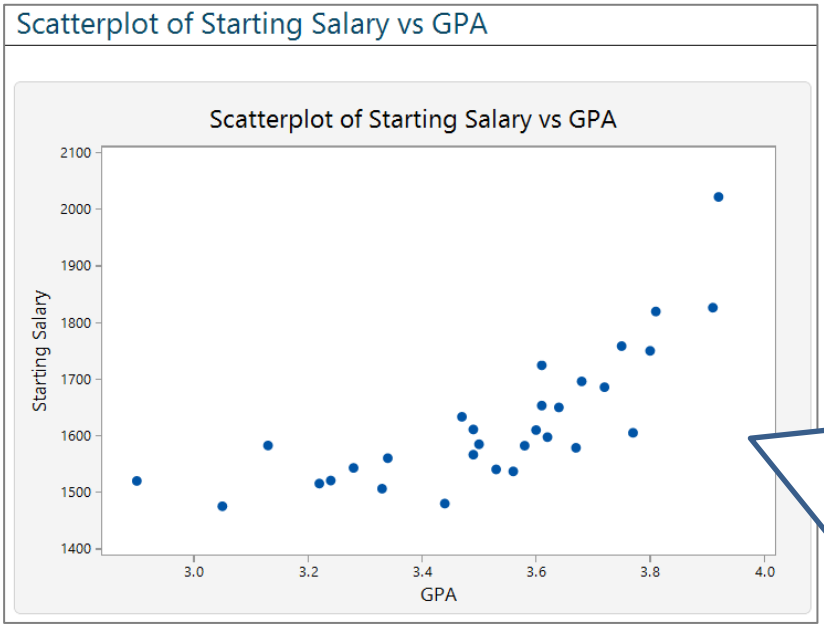
- Describe the size of your sample
  - Use "N" to know how many observations are in your sample. Minitab does not include missing values in this count.
- Describe the centre of your data
  - Use the mean to describe the sample with a single value that represents the centre of the data. Many statistical analyses use the mean as a standard measure of the centre of the distribution of the data.
  - The median and the mean both measure central tendency. But unusual values, called outliers, affect the median less than they affect the mean. When you have unusual values, you can compare the mean and the median to decide which the better measure to use. If your data are symmetric, the mean and median are similar.
- Describe the spread of your data
  - Use the standard deviation (or IQR) to determine how spread out the data are from the mean. A higher standard deviation value indicates greater spread in the data.

Task 1D

Use the dataset in "*Prac1*" Excel worksheet to construct scatterplot and compute correlation coefficient.

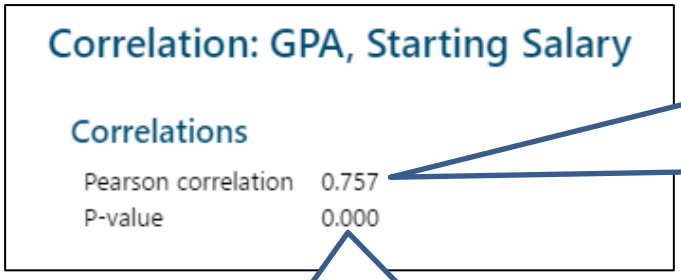
(I) Construct scatterplot for the quantitative variables *GPA* and Starting Salary.

(II) Compute correlation for the quantitative variables *GPA* and Starting Salary.



Basic Interpretations of Scatterplots

- Are the points close to an "imaginary" linear line?
- When the explanatory variable increase, does the response variable increase too or decrease?
- Hence, is this indicative of a positive or negative relationship?



"*P-Value*" of correlation coefficient will be covered in Chapter 9.

Basic Interpretations of Correlation Coefficients

- Does the value indicate a linear relationship?
- Does the value indicate a positive or negative relationship?
- Does the value indicate a strong, moderate or weak linear relationship?

Task 1E (OPTIONAL)

Use the dataset in "*Prac1*" Excel worksheet to stack and unstack data in Minitab. (Alternatively, filter data in Excel before copying data over to Minitab.)

(I) Unstack the *GPA* data according to *School*, so that we see a column of GPA for each school.

(II) Stack all the columns of *GPA* (i.e. *GPA_CLS*, *GPA_EEE*, *GPA_MAE*, *GPA_SB*) back to a single column.

Task 2

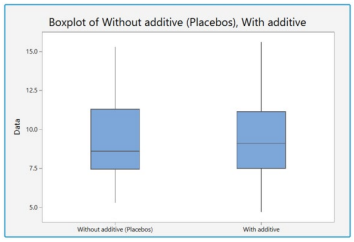
Moto Automobile's would like to know if its newly developed petrol additive is useful in increasing car mileage significantly. Fifty car owners were randomly asked to include additives into their cars, of which 25 car owners were given the petrol additives and 25 others were given placebos. All the car owners were asked to diligently and carefully record their car mileage (in km) per litre of petrol used. The results are shown as follows:

Without additive (Placebos)					With additive				
7.2	7.7	6.1	11.9	9.5	7.4	7.3	7.6	12.2	9.3
8.6	10.9	7.2	6.9	15.2	9.1	10.4	6.6	6.9	15.2
5.3	8.6	10.2	8.4	9.2	5.3	9.5	9.5	8.2	9.7
9.0	8.2	13.0	15.3	8.4	8.4	7.9	12.9	15.6	8.3
9.0	5.3	11.9	8.5	11.7	9.7	4.7	11.9	7.6	12.2

(This data set can be found in "Prac1" Excel worksheet.)

Here is the comment from one of the owners:

Is there enough evidence to support this user's comment? Justify using the data given.

Formulating Questions	
Collecting Data	Sample size, n = "Additive type" is "Mileage" is
Analysing Data	Descriptive Statistics: Without additive (Placebos), With additive Statistics Variable N Mean StDev Minimum Q1 Median Q3 Maximum Range IQR Without additive (Placebos) 25 9.3280 2.6708 5.3000 7.4500 8.6000 11.3000 15.3000 10.0000 3.8500 With additive 25 9.3360 2.7480 4.7000 7.5000 9.1000 11.1500 15.6000 10.9000 3.6500 
Interpreting Results	The numerical summaries seem to show that the average and variation in the mileage recorded by the car are _____. The boxplot generated _____ _____. Hence, there is _____ to suggest that the additive gives higher mileage per litre of petrol.

CHAPTER 2
PROBABILITY DISTRIBUTIONS

Learning Objectives:

1. *Define probability.*
2. *Determine rare events.*
3. *Distinguish between classical and empirical approaches of computing probability.*
4. *Use basic probability rules.*
5. *Define random variable.*
6. *Distinguish between discrete and continuous random variables.*
7. *Identify the Binomial random variable.*
8. *Apply the Binomial probability model.*
9. *Identify the Normal curve and its characteristics.*
10. *Find probabilities under the standard Normal curve by reading Z-table.*
11. *Convert any Normal curve to the standard Normal curve, and find the corresponding probability.*
12. *Convert any probability given from the Normal curve to find the corresponding random variable X value.*
13. *Apply the Normal distribution in application problems.*
14. *Compute the probability of the Normal distribution using Minitab.*
15. *Interpret Minitab outputs of Binomial and Normal distributions.*

Content

1. Introduction to Probability

1.1 What is Probability?

Probability is the mathematical way of quantifying _____, in order to make predictions in the real world. Some examples of chance or probabilistic statements we encounter in our daily lives are:

- Weather report says "there is a 70% chance of rain today."
- Doctor says "there is a 20% chance of complications from the surgery."
- Singapore Pools says "there is a 1 in 175 million chance of winning the lottery."
- NASA says "the probability of a giant asteroid slamming Earth is very low."

The **probability** of any event is the _____ or _____ of times it would occur in a long series of repetitions. Such chance events might be unpredictable in the short term, but has a regular and _____ in the _____.

Probabilities take values between ___ and ___, both inclusive.
Probability close to 1 means the event will _____, and probability close to 0 means the event is _____ to happen.

1.2 Calculating Probability

There are two ways to calculate probability, _____ and _____.
To illustrate both further, let us first define some terms.

- **Experiment** or **trial** is an occurrence that has an _____ outcome.
- **Sample space** is the set of _____. The probabilities of all possible outcomes will add up to ___.
- **Event** is _____ of the outcomes.
- **Rare event** is an event with _____ probability, i.e. close to _____.

Classical probability is used when each outcome in a sample space is equally likely to occur. The classic probability of an event A is given by:

$$P(A) = \frac{\text{number of outcomes in A}}{\text{total number of outcomes in sample space}}$$

Empirical probability is based on observations obtained from probability experiments. The empirical probability of an event A is simply the **relative frequency** of event A:

$$P(A) = \frac{\text{number of times A occurs}}{\text{number of times experiment is conducted}} = \frac{\text{frequency of A}}{\text{total frequency}}$$

Example 1: In each of the probability statements below, decide between classical and empirical. Also, identify the rare event.

P(getting a head in a coin toss) = 50%	Classical / Empirical
P(rain today) = 70%	Classical / Empirical
P(winning the lottery) = 0.00000001	Classical / Empirical
P(complications from surgery) = $\frac{1}{5}$	Classical / Empirical

Case Study 1: Ceramic Insulators

By Jarek Tuszyński / CC-BY-SA-3.0, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17724597

In Singapore, high-voltage electric power transmission cables are installed underground. However, in many neighbouring countries and further, these cables are above ground and overhead.

Although these cables are usually left bare and uninsulated, insulators are required at the points where they are supported by utility poles or transmission towers. Such insulators are often of ceramic material due to its non-conductivity and heat-withstanding property.

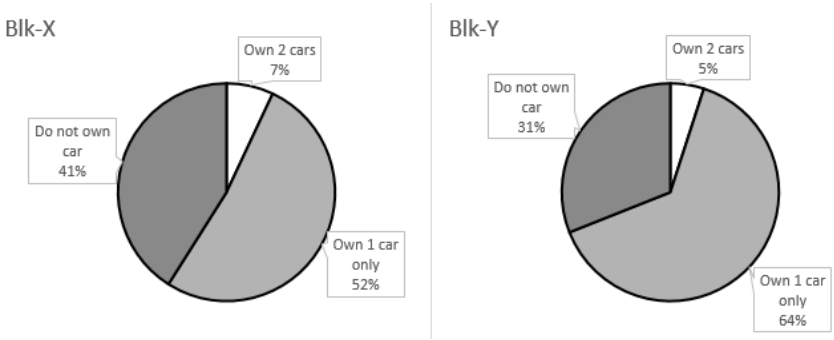
Formulating questions	As the ceramic insulators are exposed to weather elements, they are expected to be able to withstand thermal shock, which is the sudden change in temperature. What is the likelihood of a randomly selected ceramic insulator shattering due to thermal shock?
Collecting data	300 random ceramic insulators from the same manufacturer is tested, of which, 4 shattered under thermal shock.
Analyzing data	P(ceramic insulator shatter) = P(ceramic insulator withstand thermal shock) =
Interpreting results	What is the likelihood of a randomly selected ceramic insulator shattering? Is it rare for a randomly selected ceramic insulator to shatter? Should it be?

1.3 Probability Rules

To compute the probability of multiple events, we can use the following probability rules:

Addition Rule for Disjointed Events	**Multiplication Rule for Independent Events**
If event A and event B are disjointed, they do not _____.	If event A and event B are independent, they do not _____ each other.
P(A **or** B) = P(A) + P(B)	P(A **and** B) = P(A) × P(B)

Example 2: In a private estate, there are two apartment blocks, Blk-X and Blk-Y.



(a) What is the probability that a random household from Blk-X owns 1 **or** 2 cars?

(b) What is the probability that both a random household from Blk-X **and** a random household from Blk-Y do not own car?

2. Random Variables

2.1 Discrete and Continuous Random Variables

A random variable is a variable whose numeric value is based on the outcome of a random event. A random variable can be classified as discrete or continuous.

A **discrete** random variable has a _____ number of possible outcomes that can be _____. On the contrary, a **continuous** random variable has an _____ number of possible outcomes.

Example 3: Are the following random variables discrete or continuous?

Number of stocks in the Straits Times Index that have share prices increase on a given day.	Discrete / Continuous
Volume of water in a 500-ml bottle.	Discrete / Continuous
Number of highway fatalities in a country.	Discrete / Continuous
Weight of a chemical compound.	Discrete / Continuous
Room temperature at 12pm on a particular day in Singapore.	Discrete / Continuous
Number of heads that comes up when a coin is tossed four times.	Discrete / Continuous

2.2 Probability of Random Variables

Consider an experiment where a fair coin is tossed four times. Here is the sample space:

HHHH HHHT HHTH HTHH THHH HHTT HTHT THHT
HTTH TTHH THTH HTTT THTT TTHT TTTH TTTT

Remember that each of the above outcomes is _____.

Suppose that we are interested in the number of heads that comes up, so we define:

X = number of _____ in four coin tosses

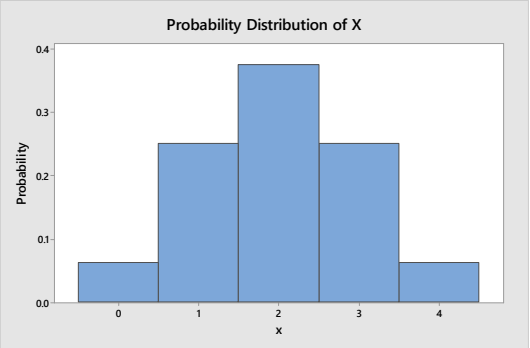
Then X is called a random variable which is the numerical outcome of a random phenomenon. Note that we do not care when in the sequence of tosses we get heads or tails, just the overall number of heads that comes up.

The **probability distribution** of a random variable X tells us the values that the random variable can take on and the probabilities of each value. Here is the probability distribution of X in our four coin tosses, presented in a table form:

Values of X					
Probability, P(X = x)					

Notice that:

- X is a _____ variable since it has a finite number of possible values.
- Each of the outcomes listed in the table is possible, but not equally likely.
- The sum of all the probabilities is ___.
- P(X = x) or p(x) denotes the probability associated with a particular value x.
- The probability distribution of four coin tosses can also be represented using a graph. The horizontal axis shows us the possible values of x, and the height of each bar represents the probability for that value.



Example 4: Use the probability distribution table of four coin tosses to answer the following questions.

(a) What is the most likely number of heads from four coin tosses?
(b) What is the probability of obtaining no heads in four coin tosses?
(c) Find P(X = 2) and P(X < 2).
(d) What is the probability of obtaining at least one head in four coin tosses?
(e) Is it rare to obtain all tails in four coin tosses?

Case Study 2: Space Shuttle Challenger

Reference: "Random Variables: Against All Odds—Inside Statistics." Films Media Group, 2013,
http://fod.infobase.com.ezp1.lib.sp.edu.sg/portalplaylists.aspx?wid=151497&xtid=111539.

On the morning of January 28, 1986, the space shuttle Challenger 7 broke apart shortly after lift-off.

After thorough investigation, a commission of experts found that the accident was caused by failure in at least one of the O-rings. The O-rings were supposed to seal field joints on the rocket boosters to contain hot, pressurized gases within the boosters.

Formulating questions

Has the risk of this failure been adequately evaluated? Could the disaster have been predicted?

Collecting data

The first step in a probability analysis of field joint failure is to calculate the probability of failure in one of them. Under the Challenger flight conditions, the probability of failure of a particular field joint is 0.023, which means that each individual field joint has a probability of success of 0.977. But a space shuttle has six field joints. So for the entire system to succeed, all six field joints have to succeed, i.e. no failures.

Analysing data

Let X be the number of failures in the six field joints.

(a) What is the probability that none of the field joints fail?

$P(X = 0) =$

(b) What is the probability that at least one field joint fail?

$P(X \geq 1) =$

Interpreting results

(a) Is the failure of a single field joint considered a rare event?

(b) Is the failure of at least one field joints considered a rare event?
What is the implication of this on the safety of a space shuttle mission?

3. Discrete Random Variable – Binomial

3.1 The Binomial Distribution

Probability models provide us with a list of all possible outcomes and proportions for how often they would each occur in the long run. We can use a probability model to find the following:

Scenario	Possible Outcomes
How many times can we expect to get heads on coin tosses?	head vs. tail
How many daffodil blossoms can we expect to see in spring, based on the number of bulbs planted in the previous autumn?	bloom vs. none
How many children in a family is expected to inherit a genetic disease?	sick vs. healthy

There is a commonality among these scenarios; they are all concerned with things that have only _____ possible outcomes. Traditionally, we think of one possible outcome as a _____ and the other as a _____. What we are interested in is the overall count of successes. The count forms a particular kind of discrete probability model, the **Binomial distribution**.

There are four conditions to identify in the Binomial distribution:

#1. There is a repeated fixed number of trials or observations, ***n***.

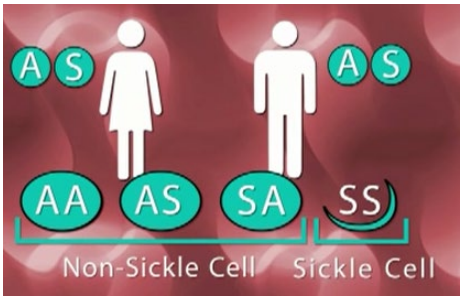
#2. All these trials are independent. That is, the outcome of one trial does not change the probabilities of other trials.

#3. Each trial should end in one of two outcomes: success or failure.

#4. The probability of success, ***p***, must be the same for all trials.

Case Study 3: Sickle Cell Disease

Reference: Binomial Distribution: Against All Odds—Inside Statistics. (2013). Films Media Group. Available at: http://fod.infobase.com/PortalPlaylists.aspx?wID=151497&xtid=111540



In people with sickle cell disease, the sickle hemoglobin molecules cause the normally round red blood cells to destore into a sickle shape, which causes blockages in the blood vessels. Tissues downstream are starved of oxygen, causing damage and much pain.

The genes that determine an individual's hemoglobin type are inherited, one version from each parent. Since it is a recessive disease, the child needs to receive two bad versions of the gene, one from each carrier parent, to have the disease.

Formulating questions	Public health officials want to know the mean number of children with sickle cell disease in a family where the parents are carriers.
Collecting data	Inheritance of the sickle cell disease, if both parents are carriers, fits the Binomial distribution. • There are 2 possible outcomes in each child conceived: sick or healthy • The outcome for each child is independent. • The number of children in a particular family and the parents' genetic makeup do not change. • The probability of a child having sickle cell disease ('success') is 0.25, and is the same for each pregnancy.
Analysing data	Let X be the number of children with sickle cell disease in a family with 6 children. Possible values of x are: 0, 1, 2, 3, 4, 5, 6 Number of trials, $n = 6$ Probability of success, $p = 0.25$ Mean, $\mu = np = 6 \times 0.25 = 1.5$
Interpreting results	So the mean number of children with sickle cell disease, in families of six children where both parents are carriers, is _____. If their first child turns out to have sickle cell disease, does it mean that their next 3 children will not have sickle cell disease? Explain.

3.2 The Binomial Probability Model

To denote a Binomial random variable X, with number of trials n and probability of success in each trial p, we write:

$$X \sim B(n, p)$$

The possible values that X can take are 0, 1, 2, 3, …, n.
Then, the probability of getting exactly x successes out of n trials is given by the formula:

$$P(X = x) = {}_nC_x \, p^x q^{n-x}$$

where q = _____ , is the probability of _____ in each trial, and

${}_nC_x$ is called the _____.

Furthermore, $\text{expected number of success} = \mu = np$

Example 5: Suppose that in a space shuttle, there are 6 field joints working independently. The probability of each field joint failing is 0.023.

(a) Explain why this scenario can fit the Binomial distribution.

(b) Let X be the number of field joints that fail out of 6. Derive the probability distribution of X.

Example 6: If two parents are both carriers of the sickle cell disease, the chance of their child having the disease is 1 out of 4.

(a) If a pair of parents with these conditions have 6 children, what is the probability that:
   (i) half of them will have sickle cell disease?
   (ii) at most one child will have sickle cell disease?

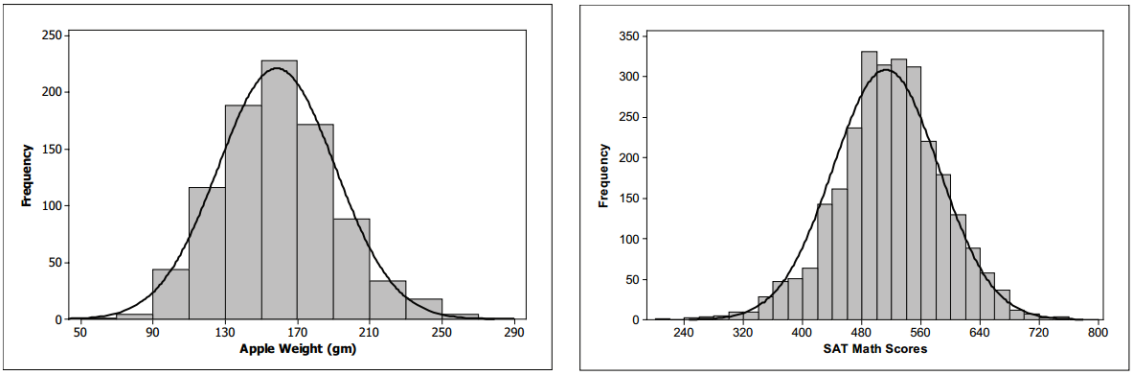
(b) Another pair of parents with the same conditions have 5 children.
   (i) How many children do they expect to have sickle cell disease?
   (ii) Is it rare to have exactly 4 children with sickle cell disease?

4. Continuous Random Variable – Normal

4.1 The Normal Curve

Histograms are often used to graph the distribution of the sample values for one particular quantitative variable. To make it even easier to focus on general shapes, sometimes statisticians draw a smooth curve through a histogram. These curves, drawn over histograms, summarize the overall patterns in data sets.

The curves can also be compared to spot similarities in shapes, even if the data sets that are being compared might not be of similar source or scale. For example, histograms of weights of Gala apples from an orchard and SAT Math scores from entering students at a US state university, are shown here:



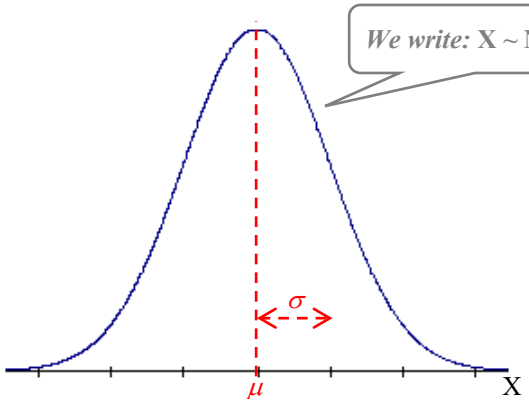
Reference: "Normal Curves: Against All Odds—Inside Statistics," director. Films Media Group, 2013, http://fod.infobase.com/portalplaylists.aspx?wid=151497&xtid=111526.

Notice that the curves on both histograms roughly have the same shape, even though the data sets are unrelated.

This special shape is called _____ curve. It is _____ with _____ peak, or simply called, **bell-shaped**. The mean μ and the median are at the same point right in the middle. In fact, many distributions in the natural world exhibit this normal curve shape.

The bars in the histogram represent the actual sample data collected. The curve represents our idealized assumption of what the whole population would look like, based on the actual data.

An important feature of any normal curve is that it can be completely defined by its **mean** μ, and its **standard deviation** σ.



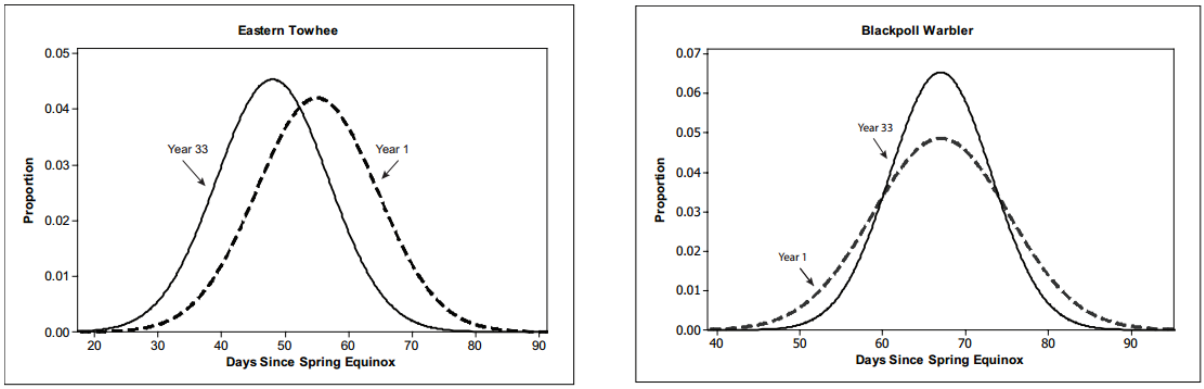
If you change the mean μ, the whole curve just shifts along the x-axis.

If you change the standard deviation σ, the distribution spread will change, such that the peak becomes flatter and wider, or becomes taller and narrower.

It is easier to make comparisons if we convert each bell-shaped smooth histogram into a **normal density curve**. To do this we change the scaling on the *y*-axis from a simple count or frequency, to a relative frequency or proportion.

With this new scale, the total area under the density curve is ___, and represents _____% of the data. Thus, 50% of the data falls below mean μ, and the other 50% of the data falls above.

Example 7: The normal density curves, in the graphs below, represent migration pattern of the Eastern Towhee and the Blackpoll Warbler birds in Manomet since 1970, at year 1 and year 33.



Reference: "Normal Curves: Against All Odds—Inside Statistics," director. Films Media Group, 2013, http://fod.infobase.com/portalplaylists.aspx?wid=151497&xtid=111526.

The following observations can be made from the Eastern Towhee graphs:

- The mean days of arrival for year 1 is later, because its curve is to the _____ of year 33's curve.
- The standard deviation of days of arrival for year ____ is smaller, because its curve is taller and pointier, and the data less spread out.
- The first arrivals for both years are happening about the same time.
- By day 48, about 50% of the birds had arrived for year 33, but only about ____% had arrived for year 1.
- About half of the birds had arrived for year 1 by day ____.

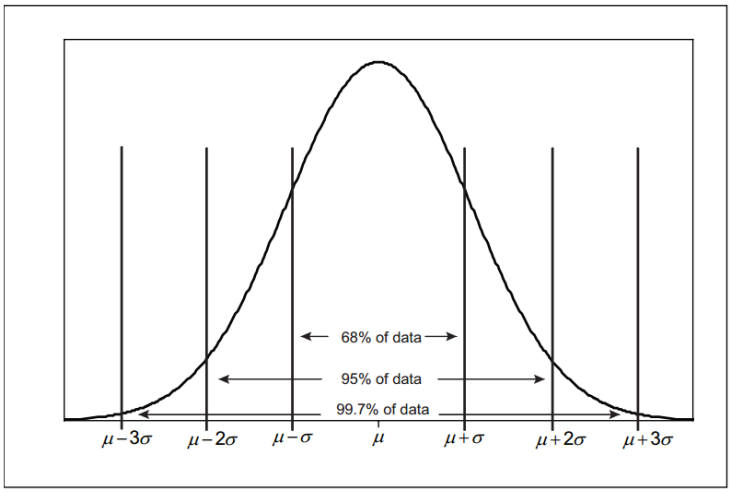
The following observations can be made from the Blackpoll Warbler graphs:

- The mean days of arrival for both years are _____.
- The first arrivals for year ____ is later.
- By day 56, about 10% of the birds had arrived for year 1, but only about ____% had arrived for year 33.
- About half of the birds had arrived for both year 1 and year 33 by day ____.

4.2 Empirical Rule

Recall that a normal curve is symmetric, single-peaked and bell-shaped, and it is completely described by its mean μ and standard deviation σ.

Normal curves have a unique feature that can be summed up by the **empirical rule**. It is also known as the _____ rule.



- Approximately _____% of the data falls within _____standard deviation of the mean.
- Approximately _____% of the data falls within _____standard deviations of the mean.
- Approximately _____% of the data falls within _____standard deviations of the mean.
- The _____ is a natural yardstick for any measurements that follow a normal distribution.

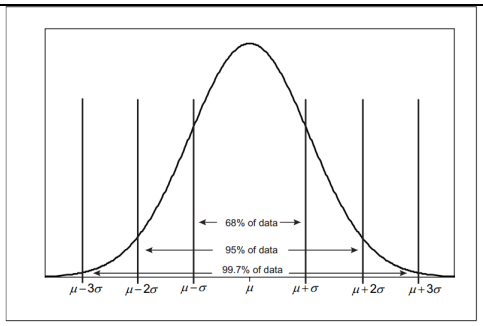
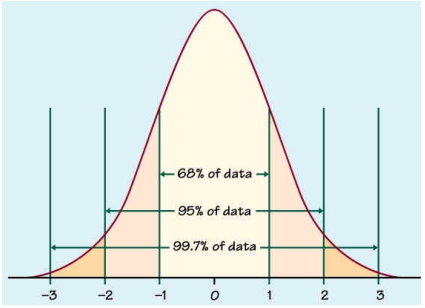
Example 8: The birth weight of (full-term) babies in the United States (U.S.) is normally distributed with mean $\mu = 3.4$ kg and standard deviation $\sigma = 0.5$ kg.
Sketch the normal curve for the birth weight of U.S. babies, marking out the standard deviation as the "yardsticks".

Hence,

- approximately _____% of U.S. babies weigh between _____ kg and _____ kg;
- approximately _____% of U.S. babies weigh between _____ kg and _____ kg;
- approximately _____% of U.S. babies weigh between _____ kg and _____ kg;
- approximately _____% of U.S. babies weigh below _____ kg or above _____ kg.

4.3 Z-Score

We can figure out what is called the **standardized** value of any observation. This unitless value, often called a **z-score**, tells us how many _____ our observation falls from the mean and in which direction. It is a way to convert data from a normal distribution into a **standard normal distribution**. Let's see the similarities and differences:

normal distribution	**standard normal distribution**
	
Centre is at mean μ	Centre is at mean $\mu = 0$
Spread is standard deviation σ	Spread is standard deviation $\sigma = 1$
We write: $X \sim N(\mu, \sigma^2)$	We write: $Z \sim N(0, 1^2)$
Empirical rule: • about 68% of data between $\mu-\sigma$ and $\mu+\sigma$ • about 95% of data between $\mu-2\sigma$ and $\mu+2\sigma$ • about 99.7% of data between $\mu-3\sigma$ and $\mu+3\sigma$	Empirical rule: • about 68% of data between −1 and 1 • about 95% of data between −2 and 2 • about 99.7% of data between −3 and 3

Mathematically, to convert x-value to z-score: $\boxed{Z = \dfrac{X - \mu}{\sigma}}$

Z-scores allow us to compare observations from two different normal distributions by standardizing them with a common scale.

Example 9: The birth weight of (full-term) babies in the United States (U.S.) is normally distributed with mean $\mu = 3.40$ kg and standard deviation $\sigma = 0.50$ kg. Comparatively, the birth weight of (full-term) babies in Germany is normally distributed with mean $\mu = 3.56$ kg and standard deviation $\sigma = 0.45$ kg. A guideline is to be set such that babies who weigh 4.5 kg and above at birth is considered "overweight" and will require closer monitoring by pediatricians.

(a) Convert the guideline to z-scores for both countries.

(b) Sketch the standard normal curve and locate the z-scores on the curve.

(c) Which of the two countries will have more "overweight" babies?

4.4 Reading Probability from Z-Table

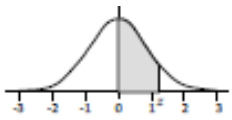
How do we find the proportion of data if the z-score does not fall exactly at 1, 2 or 3 standard deviations away from the mean?

The area under the standard normal density curve represents the proportion of data or probability. This can be found using a standard normal table or statistical software.

There are different formats of the standard normal table (or simply, z-table). We shall use the format that shows areas (probabilities) that are measured from the centre of the standard normal curve, that is, from zero to the desired z-score.

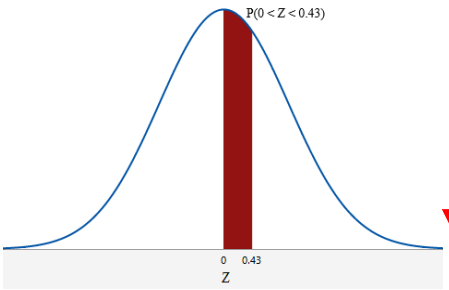
Area under the Standard Normal Curve from 0 to z

$$z = \frac{x - \mu}{\sigma}$$



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

So, how do we read this z-table? Here is an illustration:



Given z-score = 0.43, then
$P(0 < Z < 0.43) = 0.1664$

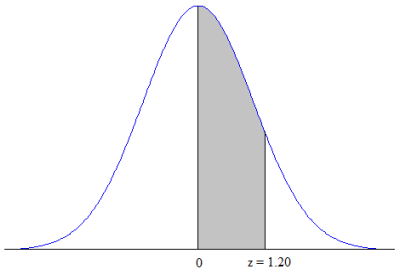
Read down the first column for 1st decimal place of z-score, then the first row for the 2nd decimal place of z-score.

z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879

Reading from the z-table, the area between 0 and desired z-score 0.43 is 0.1664, or 16.64%.

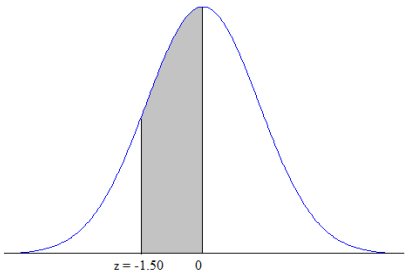
Example 10: Find the probability in each of the following:

(a)



$P(0 < Z < 1.20) =$

(b)

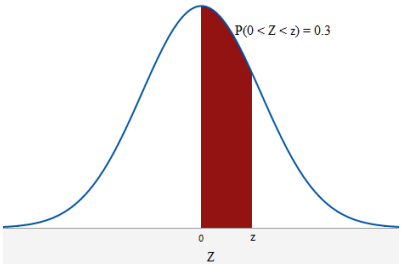


$P(-1.50 < Z < 0) =$

(c) $P(Z < 0.37) =$

(d) $P(-0.72 < Z < 1.5) =$

Conversely, if we are given the probability (area), how do we get the corresponding z-score from the z-table? Here are two illustrations:



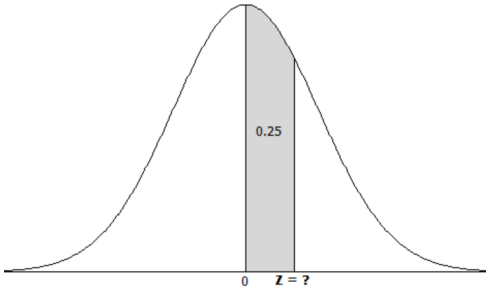
Given area between 0 and z-score is 0.3, then z = 0.84.

z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621

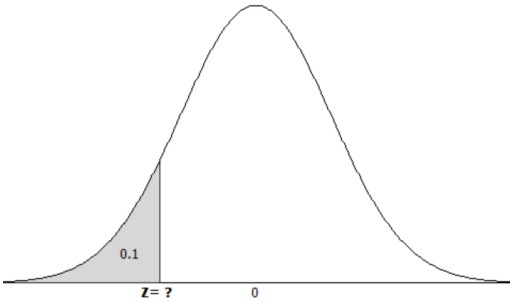
Reading from z-table, area of 0.3 (or 30%) is closest to area of 0.2996. Tracing back to the 1st column and 1st row, gives desired z-score of 0.84.

Example 11: Find the unknown values in each of the following:

(a)



(b)



(c) Given P(Z > a) = 0.2676, find a.

(d) Given P(Z > c) = 0.6845, find c.

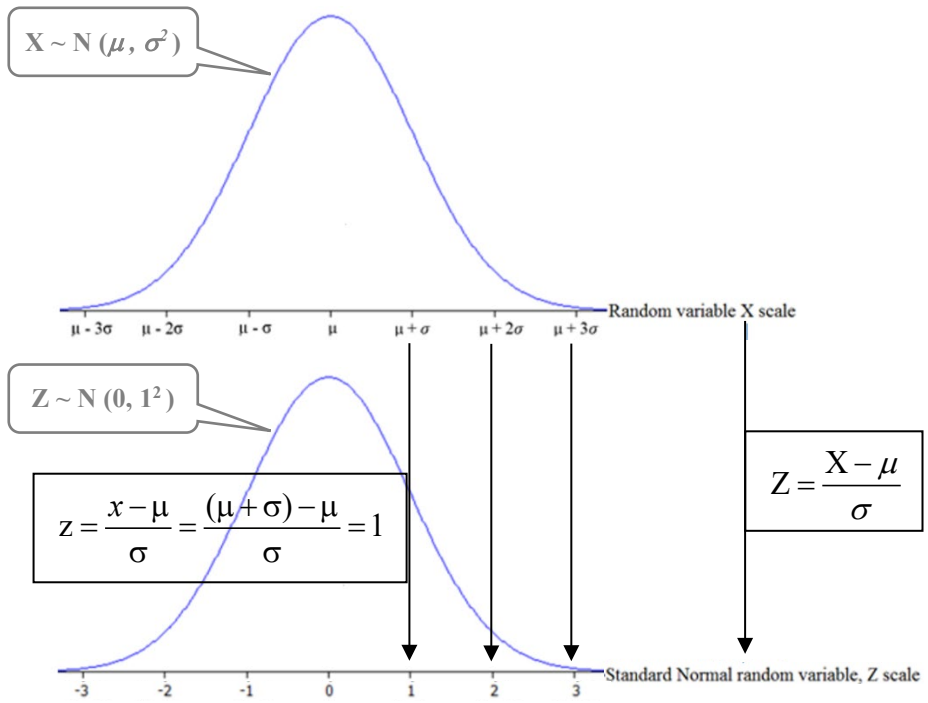
(e) If P(–1.7 < Z < m) = 0.0666, find m.

(f) If P(–n < Z < n) = 0.251, find n.

4.5 Normal Random Variable X

We now know how to obtain probabilities related to z-scores, not restricted by the Empirical Rule. What if we want to find probabilities concerned with any normal random variable X?

Any normal curve can be changed into the **standard** normal curve by transforming the scale on the horizontal axis into a z-score scale.



Example 12: Given that a random variable X has a normal distribution with $\mu = 50$ and $\sigma = 10$, find the following probabilities:

(a) $P(X \geq 45)$

(b) $P(X > 62)$

(c) $P(45 < X < 62)$

Example 13: Back to "overweight" babies. Let random variable X be the birth weight (in kg) of a U.S. baby, so $X \sim N(3.40, 0.50^2)$. Let random variable Y be the birth weight (in kg) of a German baby, so $Y \sim N(3.56, 0.45^2)$. Recall that the guideline of classifying "overweight" babies is 4.5 kg and above.

(a) Find the proportion of U.S. babies classified as "overweight".

(b) Find the proportion of German babies classified as "overweight".

(c) Find the percentage of German babies who weigh above 4 kg at birth, but below guideline of "overweight".

(d) What is the birth weight of a U.S. baby who is in the 90^{th} percentile?

Example 14: The measurements inside a diameter of a cast-iron pipe is normally distributed with mean 5.01 cm and standard deviation 0.025 cm. The specification limits are set at 5.00 ± 0.05 cm. What percentage of the pipes are not acceptable?

Example 15: The dimension of a circular mechanical part is normally distributed with mean 2 cm. The specification limits are set at 2.00 ± 0.05 cm.

(a) If the standard deviation is 0.03 cm, what is the percentage of mechanical parts that are within the specification limits?

(b) How small must the standard deviation be if 95% of the mechanical parts must be within specification limits?

TUTORIAL 2

1. A box contains 100 items, of which, 27 are oversized and 16 are undersized. The items which are not of the right size will be rejected and the rest will be accepted.
   (a) What is the probability that a randomly selected item from this box is undersized?
   (b) What is the probability that a randomly selected item from this box is accepted?

2. In a class of 25 students who took a Mathematics test, the grade distribution is as follows:

Grade	A	B	C	D	Fail
Count	8	10	4	2	1

   (a) What is the probability that a randomly selected student from this class scored 'A'?
   (b) What is the failure rate (i.e. percentage of students who failed) of this class?
   (c) What is the probability that a randomly selected student from this class scored 'C' or 'D'?

3. In a sample of 446 cars stopped at a roadblock, 34 of the drivers did not have their seatbelts fastened. Among these 34 drivers, 21 of them were first-time offenders and let off with a warning; the rest were given demerit points and fined.
   (a) What is the sample space in this scenario?
   (b) What is the probability that a random driver stopped at that roadblock will have his/her seatbelt fastened?
   (c) What is the probability that a random driver stopped at that roadblock will be given demerit points and fined for not fastening seatbelt? Is this a rare event?

4. A random sample of 250 youths between 18 and 25 years-old was selected and asked about the number of email accounts they signed up for and their primary email account which they use most often. The data collected is tallied into the following table:

		Number of email accounts			
		1	2	3	4 or more
Primary email account	Gmail	30	28	17	7
	Outlook	25	31	26	10
	Yahoo	20	26	19	11

   If a random youth is selected, what is the probability that the youth:
   (a) has 2 email accounts and uses Gmail primarily
   (b) has and uses Outlook only (i.e. has only 1 email account)
   (c) has 1 email account only
   (d) uses Yahoo primarily
   (e) does not use Gmail primarily
   (f) has at least 2 email accounts
   (g) has at least 2 email accounts and uses Outlook primarily
   (h) has at most 2 email accounts and does not use Yahoo primarily

5. Decide whether each of the scenarios described below fits the Binomial distribution.
   If it does, identify the values of n, p and q, and list all the possible values of X.
   If it does not, explain why.
   (a) Cyanosis is the condition of having bluish skin due to insufficient oxygen in the blood. About 80% of the babies born with cyanosis recover fully. A hospital is caring for five babies born with cyanosis. The random variable X represents the number of babies that fully recover from cyanosis.
   (b) A survey company called 1000 people to ask whether they "agree, disagree or have no opinion" about the latest suggestion to abolish national service in a certain country. The random variable X represents the number of people in the survey who agree to the suggestion.
   (c) An inventory study determines that, on average, demand for a certain type of item is made 5 times a day. The random variable X represents the number of demands for that item per day.
   (d) It is conjectured that an impurity exists in three out of ten drinking wells in a rural community. To gain some insights to the extent of the problem, it is determined that some testing is necessary. Since testing all the wells is too resource-intensive, 15 wells were randomly selected for testing. The random variable X is the number of wells that contain impurity.

6. Refer to the coin toss example in Section 2.2 on page 5, where X is the number of heads that comes up in four coin tosses.
   (a) Explain why X can be a Binomial random variable.
   (b) What is the mean number of heads tossed in four coin tosses?
   (c) Use the Binomial probability model to derive the probability distribution of X.

7. Given $X \sim B(12, 0.4)$, find the following probabilities:
   (a) $P(X = 2)$ (b) $P(X < 3)$ (c) $P(X \geq 4)$ (d) $P(2 \leq X < 5)$

8. Of all the students in a school, 30% travel to school by school bus. In a random sample of 10 students selected, find the following:
   (a) What is the mean number of students who travel to school by bus?
   (b) What is the probability that exactly 3 students travel to school by bus?
   (c) What is the probability that half of the students travel to school by bus?
   (d) Is it rare that more than 8 students travel to school by bus?

9. A space shuttle has two boosters, with each booster having three field joints. The field joint design has been improved such that the success launch rate of each field joint is now 0.985 (instead of 0.977). Let X be the number of failures in the six field joints, and assume that the six field joints are independent.
   (a) What is the probability that a single field joint will fail?
   (b) What is the probability that exactly one field joint in the shuttle will fail?
   (c) What is the probability that at least one field joint in the shuttle will fail?
   (d) Has the safety of a space shuttle mission improved?
   (e) Does X still fit a Binomial distribution if the field joints are not independent of each other?

10. An inspection plan for a microchip factory operates as follows:

    A sample of ten microchips are randomly selected from a large batch.
    If none from the sample is defective, then accept the batch.
    If more than one microchip from the sample is defective, then reject the batch.
    If exactly one microchip from the sample is defective, take another sample of ten microchips from the batch, and accept the batch only if none from this second sample is defective.

    If a batch of microchips with 5% defectives is inspected, find the following:

    (a) What is the mean number of defectives in the sample?
    (b) What is the probability that the batch is accepted after the first sampling?
    (c) What is the probability that the batch is accepted only after the second sampling?
    (d) What is the probability that the batch is accepted?

11. Answer the following questions based on the Minitab output given for a Binomial random variable:

    (a) How many trials were conducted?
    (b) What is the probability of a success? Is this a rare event?
    (c) What is the probability of four successes? Is this a rare event?

Probability Density Function

Binomial with n = 35 and p = 0.5

x	P(X = x)
4	0.0000015

12. Answer the following questions based on the Minitab output given for a Binomial random variable:

    (a) How many trials were conducted?
    (b) What is the probability of a success?
    (c) What is the probability of getting more than 20 successes?
    (d) What is the probability of getting at least 20 successes?

Cumulative Distribution Function

Binomial with n = 40 and p = 0.35

x	P(X ≤ x)
20	0.982719

13. Suppose that the random variable X is normally distributed with mean $\mu = 86$ and standard deviation $\sigma = 5$.

    (a) Sketch the normal curve and apply the Empirical Rule.
    (b) Hence, use the Empirical Rule to estimate the following probabilities:
        (i) $P(X < 96)$ (ii) $P(X \le 81)$ (iii) $P(76 \le X < 91)$
    (c) Convert the indicated x-value to z-score, hence find the corresponding probabilities.
        (i) $x = 80$; $P(X < 80)$ (ii) $x = 92$; $P(X > 92)$
        (iii) $x = 100$; $P(X < 100)$ (iv) $x = 72$; $P(X > 72)$
        (v) $x = 70$; $P(70 < X < 80)$ (vi) $x_1 = 85, x_2 = 95$; $P(85 < X < 95)$

14. Find the desired z-scores given the following probabilities:

    (a) $P(0 < Z < a) = 0.4753$ (b) $P(b < Z < 0) = 0.129$

    (c) $P(Z < c) = 0.97$ (d) $P(Z > d) = 0.864$

    (e) $P(Z > k) = 0.0217$ (f) $P(Z < l) = 0.271$

    (g) $P(-1 < Z < m) = 0.5$ (h) $P(1.5 < Z < w) = 0.0018$

15. Telephone calls from a call centre are monitored and found to have a mean duration of 452 seconds and a standard deviation of 123 seconds. Suppose that the distribution of call durations is approximately normal, determine the following:

    (a) What is the percentage of calls that last more than 10 minutes (i.e. 600 seconds)?

    (b) What is the percentage of calls that last more than 5 minutes?

    (c) What is the percentage of calls with duration between 300 seconds and 480 seconds?

    (d) If 250 calls are made from the call centre on a particular day, what is the mean number of calls that last more than 5 minutes? Round off your answer to the nearest whole number.

16. Components made by a certain process have a thickness which is normally distributed about a mean of 3.00 cm and a standard deviation of 0.03 cm. A component is classified as defective if its thickness lies outside the limits of 2.95 cm to 3.05 cm. Find the proportion of defective components. Hence, in a batch of 500 components, how many will be defective on average?

17. The mass of a bag of cookies is normally distributed with mean 450 g and standard deviation 15 g. Bags of cookies that have mass in the upper 7.5% are too heavy and must be repackaged. What is the maximum allowable mass a bag of cookies can be such that repackaging is not required?

18. You sell a brand of automobile tyre which has a life expectancy that is normally distributed with a mean of 30,000 km and a standard deviation of 2,500 km. You want to give a guarantee for free replacement of tyres that do not wear well. How should you word your guarantee if you are willing to replace approximately 10% of the tyres you sell?

19. A normal distribution has mean $\mu = 62.4$. Find its standard deviation if 20% of the area under the curve lies to the right of 79.2.

20. A vending machine is calibrated to dispense coffee into a 250ml paper cup. The amount of coffee dispensed into the cup is normally distributed with a standard deviation of 10ml. If the machine is allowed to overfill the cup 1% of the time, what should be set as the mean amount of coffee to be dispensed?

21. Suppose the life in hours of a certain electronic tube is normally distributed with mean $\mu = 160$ hours. The specification limits call for the product to last between 120 hours and 200 hours with probability 0.95. What is the maximum allowable standard deviation that the process can have and still maintain its quality?

ANSWERS[1]

1. (a) 0.16 (b) 0.57
2. (a) 0.32 (b) 0.04 (c) 0.24
3. (a) "fastened seatbelt", "did not fasten seatbelt and let off with warning", "did not fasten seatbelt and given demerit points and fined" (b) 0.924 (c) 0.0291, yes
4. (a) 0.112 (b) 0.1 (c) 0.3 (d) 0.304 (e) 0.672 (f) 0.7 (g) 0.268 (h) 0.456
5. (a) Yes. $X \sim B(n = 5, p = 0.8)$, $q = 0.2$
   (b) No, because there are 3 possible outcomes in each trial – agree, disagree, neutral. However, if we group the outcomes such that 'success' is agree and 'failure' is either disagree or neutral, then we can fit a Binomial distribution $X \sim B(n = 1000, p = 1/3)$.
   (c) No, because the number of trials is not fixed, but possibly infinite.
   (d) Yes. $X \sim B(n = 15, p = 0.3)$, $q = 0.7$
6. (a) Fixed number of trials (four tosses => $n = 4$); only 2 possible outcomes (H or T); probability of success is constant from trial to trial ($p = 0.5$), this means the trials are independent. (b) 2 (c) Refer to probability distribution table on page 5.
7. (a) 0.0639 (b) 0.0834 (c) 0.775 (d) 0.419
8. (a) 3 (b) 0.267 (c) 0.103 (d) $P(X > 8) = 0.000144$, rare
9. (a) 0.015 (b) 0.0834 or [0.0835] (c) 0.0867 (d) Improved (was 13%), but still not safe because failed mission is not a rare event. (e) No, because it will violate one of the conditions of a Binomial distribution.
10. Let X be the number of defective microchips in a sample, then $X \sim B(10, 0.05)$
    (a) 0.5 (b) 0.5987 (c) 0.1887 (d) 0.7874
11. (a) 35 (b) 0.5, no (c) 0.0000015, yes
12. (a) 40 (b) 0.35 (c) 0.0173 (d) 0.0363
13. (a) About 68% between 81 and 91, about 95% between 76 and 96, and about 99.7% between 71 and 101. (b) (i) 97.5% (ii) 16% (iii) 81.5%
    (c) (i) –1.2, 0.1151 (ii) 1.2, 0.1151 (iii) 2.8, 0.9974
    (iv) –2.8, 0.9974 (v) –3.2, 0.1144 (vi) –0.2, 1.8, 0.5434
14. (a) 1.965 (b) –0.33 (c) 1.88 (d) –1.10 (e) 2.02 (f) –0.61 (g) 0.41 (h) 1.51
15. (a) 11.51% or [11.44%] (b) 89.25% or [89.17%] (c) 48.35% or [48.18%] (d) 223
16. 0.095, 47.5 or [0.096, 48] 17. 471.6 g
18. "Tyres that wear out before 26,800 km [or 26,796 km] will be replaced free of charge!"
19. 20 20. 226.7 ml 21. 20.41 hours

[1] Answers obtained using Minitab may sometimes be slightly different from answers obtained using *z*-table. For such cases, answers from Minitab are in square brackets [ ].

PRACTICAL 2A : Binomial Distribution

Learning Objectives:

1. *Find probability of Binomial distribution using Minitab.*
2. *Find cumulative probability of Binomial distribution using Minitab.*

Task 1

Consider an experiment of tossing a dice 30 times. Assuming that the dice is fair, what is the probability of not getting any '6' in 30 tosses of a dice?

Let X represent the Binomial random variable: number of '6' in 30 tosses of a dice.

So, $X \sim B\left(n = 30,\ p = \frac{1}{6}\right)$

Using the Binomial formula, $P(X = 0) = {}_{30}C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{30} = 0.0042$. Now, use Minitab to obtain the same answer.

Task 2

Assuming that the dice is fair, what is the probability of getting at most five '6' in 30 tosses of a dice?

For this, calculation using Binomial formula is possible, but tedious!

$P(X \le 5) = P(X = 5) + P(X = 4) + P(X = 3) + P(X = 2) + P(X = 1) + P(X = 0) = 0.62$

Using Minitab to obtain this will be much easier!

PRACTICAL 2B : Normal Distribution

Learning Objectives:

1. *Find probability of standard normal distribution using Minitab.*
2. *Find probability of any normal distribution using Minitab.*
3. *Find the z-score given the probability in the standard normal distribution using Minitab.*

Task 1

On page 15 of this chapter, we found that P(0 < Z < 0.43) = 0.1664 by using the z-table. Let us obtain this probability using Minitab.

Task 2

Suppose X is normally distributed with $\mu = 100$ and $\sigma = 20$, let us obtain P(90 < X < 140) using Minitab.

Task 3

Suppose we know that P(Z > *k*) = 0.1314, that is, the area under the standard normal curve to the right of *k* is 0.1314, let us use Minitab to find the value of *k*.

(Optional) Investigative Task

In Minitab, plot the Binomial distribution for *p* = 0.3, with different values of *n* from small (say, 10) to very large (say, 1000). What do you observe about the shape of the distribution as *n* gets larger? (You may repeat this for another value of *p*.)

CHAPTER 3A
SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Learning Objectives:

1. *Define sampling distribution of the sample mean, $\bar{X}$.*
2. *Link the mean and standard deviation of sampling distribution of the sample mean, $\bar{X}$ to the mean and standard deviation of random variable X.*
3. *Understand the Central Limit Theorem and its role in statistical inference.*
4. *Use sampling distributions to evaluate claims ("rare-events") on values of the population parameter μ.*

Content

1. Case Study: Subway

Subway 'crisis': Is footlong sub really 11 inches?

NEW YORK (AP) — What's in an inch? Apparently, enough missing meat, cheese and tomatoes to cause an uproar.

Subway, the world's largest fast food chain with 38,000 locations, is facing widespread criticism after a man who appears to be from Australia posted a photo on the company's Facebook page of one of its foot-long sandwiches next to a tape measure that shows the sub is just 11 inches.

More than 100,000 people have "liked" or commented on the photo, which had the caption "Subway pls respond." Lookalike pictures popped up elsewhere on Facebook. And The New York Post conducted its own investigation that found that four out of seven foot-long sandwiches that it measured were shy of the 12 inches that makes a foot.

The original photo was no longer visible by Thursday afternoon on Subway's Facebook page, which has 19.8 million fans. A spokesman for Subway, which is based in Milford, Conn., said Subway did not remove it.

Subway also said that the length of its sandwiches may vary slightly when its bread, which is baked at each Subway location, is not made to the chain's exact specifications. "We are reinforcing our policies and procedures in an effort to ensure our offerings are always consistent no matter which Subway restaurant you visit," read an e-mailed statement.

The Subway photo — and the backlash — illustrates a challenge that companies face with the growth of social media sites like Facebook, YouTube and Twitter. Before, someone in a far flung local in Australia would not be able to cause such a stir. But the power of social media means that negative posts about a company can spread from around the world in seconds.

"People look for the gap between what companies say and what they give, and when they find the gap — be it a mile or an inch — they can now raise a flag and say, 'Hey look at this,' I caught you," said Allen Adamson, managing director of branding firm Landor Associates in New York.

Subway has always offered foot-long sandwiches since it opened in 1965. A customer can order any sandwich as a foot-long. The chain introduced a $5 foot-long promotion in 2008 as the U.S. fell into the recession, and has continued offering the popular option throughout the recovery.

An attempt to contact someone with the same name and country as the person who posted the photo of the foot-long sandwich on Subway's Facebook page was not returned on Thursday.

But comments by other Facebook users about the photo ran the gamut from outrage to indifference to amusement. One commenter urged people to "chill out." Another one said she was switching to Quiznos. And one man posted a photo of his foot in a sock next to a Subway sandwich to show it was shorter than a "foot."

"I've never seen so many people in an uproar over an inch. Wow," read one Facebook post. "Let's all head to McDonald's and weigh a Quarter Pounder," suggested another poster.

*(Retrieved from: **http://tinyurl.com/yahoonews-subway**)*

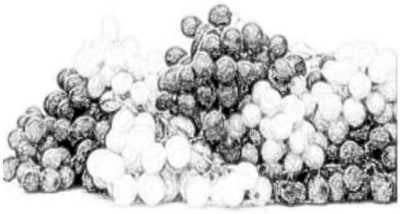
Upon reading the news article, Thomas bought 30 foot-long subs and measured each of them. He recorded the readings as follows:

12.0	11.5	11.9	12.1	12.0	11.8	12.2	12.4	12.0	11.9
12.6	12.0	11.6	12.7	12.1	12.0	11.4	11.8	11.7	12.4
11.6	11.9	11.7	11.4	11.6	11.9	11.6	12.1	11.8	11.6

Based on Thomas's measurements, does he have enough evidence to believe that Subway subs are less than 12 inches?

2. Sampling Techniques

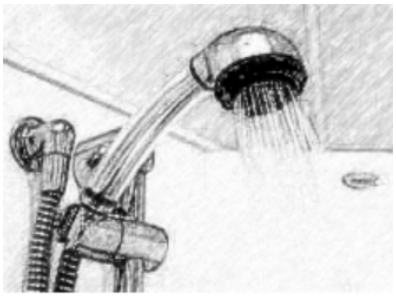
What is sampling? Observe two scenarios below:



"Hmm…I wonder if these grapes are sweet"

"Let me try one grape…Hmm, it is sweet!"

"Ok, I will buy this bunch of grapes then…"



"Is the water too hot?"

"Let me feel the water quickly… hmm, just nice…not too hot"

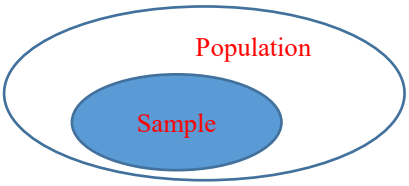
"Ok, the water temperature is just nice to shower this puppy"

What do you observe in both scenarios? Hopefully we observed the same thing:

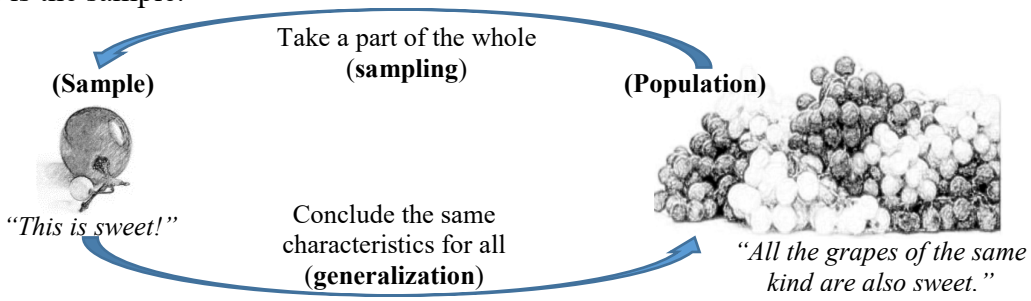
- In the first, the lady tasted one grape and concluded that all the grapes taste sweet.
- In the second, the puppy owner tested the temperature of the shower water for a brief moment and concluded that the temperature of the running water will not be too hot.

What we observe in both scenarios are examples of sampling.

Sampling is taking a part (sample) of the whole (population). So, a sample is a set of observations that is part of all the possible observations of a phenomenon.



For example, all the grapes of the same kind by the seller is the population and the one grape tasted is the sample.

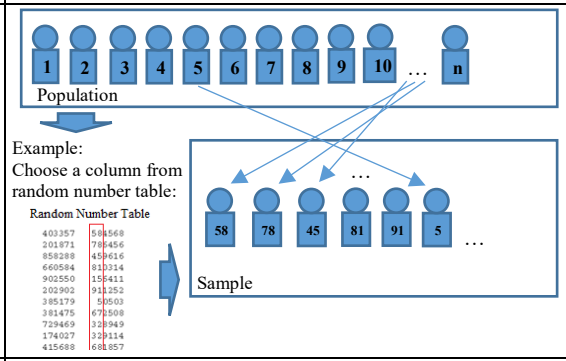
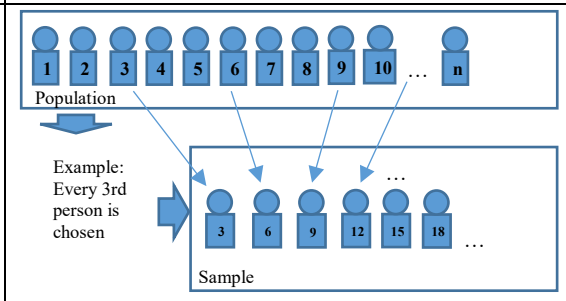
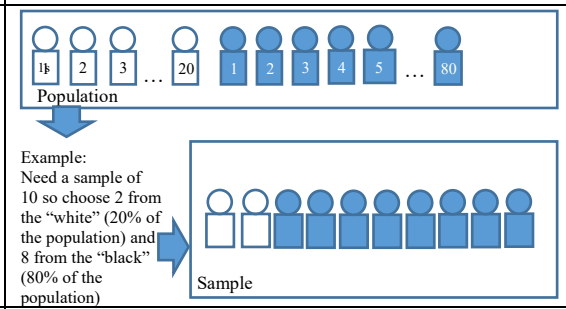
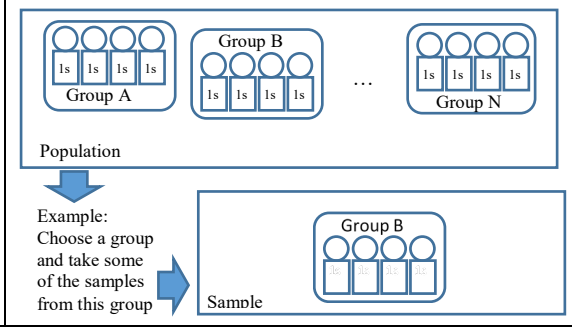


The field of statistics is ultimately concerned with generalization and prediction. In many cases, sampling is more feasible to study than the entire population.
Why? Well, the lady cannot possibly taste all the grapes before she buys nor can the dog owner let the water run forever.

Generally, here are three possible reasons why sampling is more feasible, with an example cited for each reason.

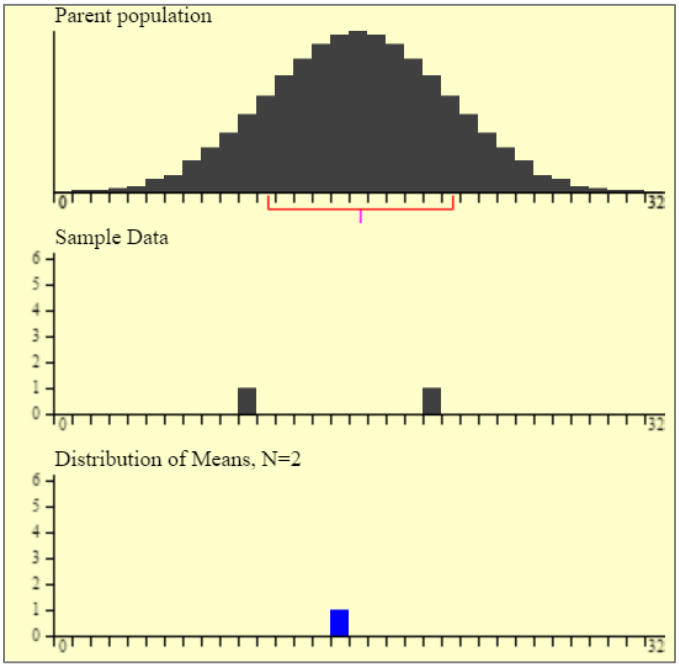
Possible Reason	Example
It is impossible to take all observations from an infinite population.	Amount of salt in the sea water in South China Sea.
Sampled objects for observation may not be returned to the population.	Impact testing on cars.
Do not have the resource (time and money) to collect all observations.	Everyone who is eligible to vote in the 2017 U.S. election and who would they vote for, candidate A or B?

There are different techniques to perform sampling; four of them are described below.

Sampling techniques	Illustration
Simple random sampling A sample is selected in such a way that every item or person in the population has the same chance of being selected.	
Systematic sampling A random starting point is selected and then every k^{th} member of the population is selected.	
Stratified sampling A population is divided into subgroups and a sample is selected from all subgroups according to the proportions of each subgroup (proportion of the population).	
Cluster sampling A population is divided into clusters or groups. Then clusters are randomly selected and a sample is collected by randomly selecting from each cluster.	

3. Distribution of Sample Means

We shall use an *applet** to demonstrate what the probability distribution of samples means look like if we are to repeatedly draw samples from the population.
[*link to applet: http://onlinestatbook.com/stat_sim/sampling_dist/index.html]

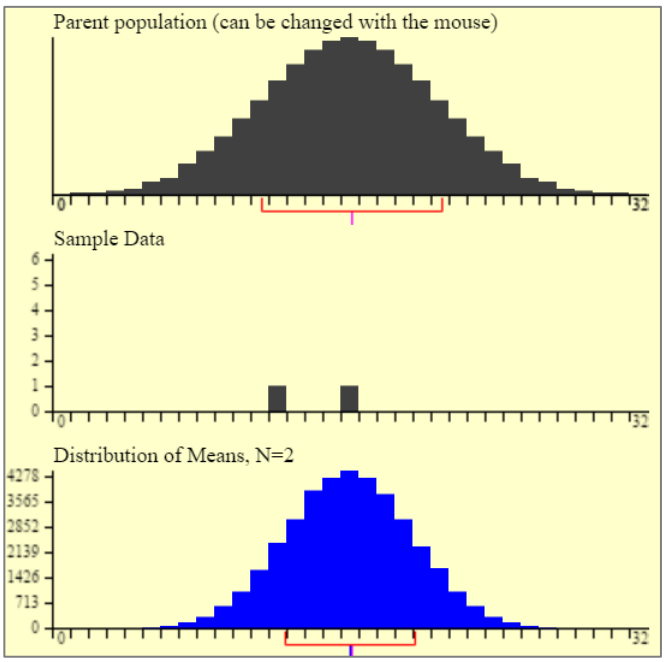


❶ Population data X is normally distributed. i.e. $X \sim N(\mu, \sigma^2)$.

❷ Two data x_1 and x_2 is randomly drawn from the population.
i.e. sample size, $n = 2$

❸ The mean of the two data x_1 and x_2 is computed, that is the sample mean $\bar{x}$.

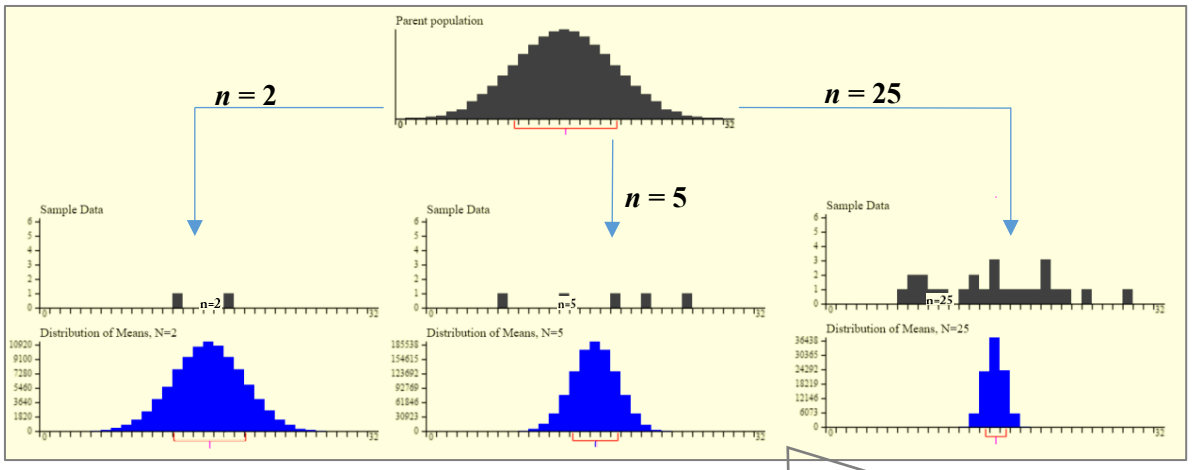
What if we repeatedly do steps ❶ and ❷ above? That is, what if we repeatedly draw 2 sample data and compute the sample mean? Surely, we will have many sample means. If we plot all these sample means on a graph, how will the distribution look like?



So, if we repeatedly draw samples of size 2 ($n = 2$) from a normal population, compute the sample means $\bar{x}$ and plot them on a graph, we will observe that the distribution of the sample means is bell-shaped!

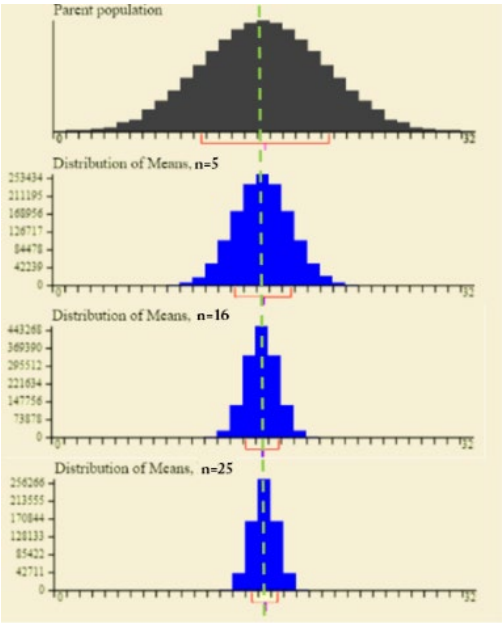
This distribution of sample means or simply, distribution of $\bar{X}$, is also known as the **sampling distribution of the sample mean**.

In fact, the sampling distribution of the sample mean is always normally distributed for any sample size n if the population is normally distributed, as illustrated below:



Think-out-loud…
What do you notice about the spread of the sampling distributions as n increases?

The distribution of the population, together with the sampling distributions of the sample mean for $n = 5$, $n = 16$ and $n = 25$ are shown here for comparisons:



Population X is normal,
centre at mean μ, spread σ.

Distribution of $\bar{X}$ of sample size $n = 5$,
centre at mean $\mu_{\bar{X}}$, spread $\sigma_{\bar{X}}$.

Distribution of $\bar{X}$ of sample size $n = 16$,
centre at mean $\mu_{\bar{X}}$, spread $\sigma_{\bar{X}}$.

Distribution of $\bar{X}$ of sample size $n = 25$,
centre at mean $\mu_{\bar{X}}$, spread $\sigma_{\bar{X}}$.

Observe that:

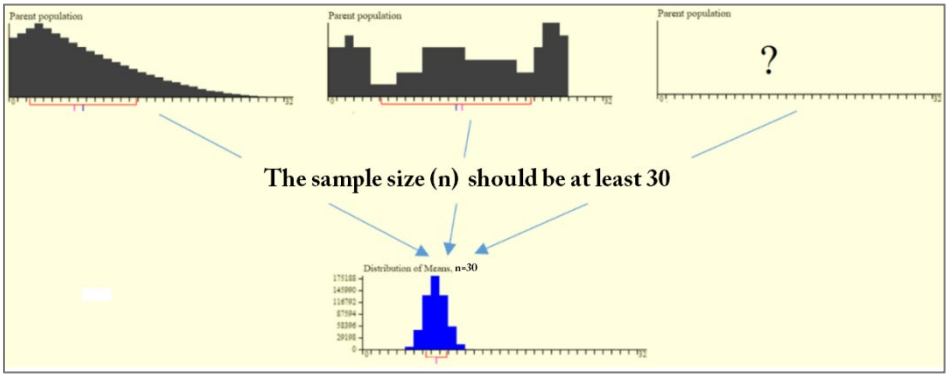
#1. The centre of the sampling distribution of the sample mean is the same as the centre of the population. That is, $\mu_{\bar{X}} = \mu$.

#2. The spread of the sampling distribution of the sample mean is smaller than the spread of the population. That is, $\sigma_{\bar{X}} < \sigma$, for $n \geq 2$.

#3. Moreover, the spread of the sampling distribution the sample mean decreases as the sample size increases. In fact, $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$.

What if the population X is not normally distributed or its distribution is unknown?

The **Central Limit Theorem** (CLT) roughly states that as the sample size *n* increases, the sampling distribution of the sample mean will approach the normal distribution.



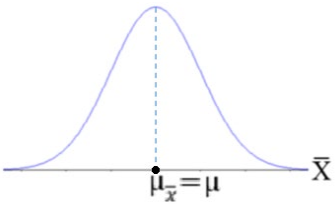
But how large should *n* be? The rule of thumb is that sample size of at least 30 is sufficient to assume that the sampling distribution of the sample mean is approximately normal.

In summary…

The sampling distribution of the sample mean is normally distributed if

- the samples are drawn from a normal population, or
- $n \geq 30$ so that CLT can be applied when the population distribution is unknown or not normal.

Since the sampling distribution of the sample mean is normal, we have that:



- Mean of sample means, $\mu_{\bar{X}} = \mu$
- Standard deviation of sample means, $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$
- $\sigma_{\bar{X}}$ is better known as **standard error** (SE) of sample mean.
- We can write as: $\bar{X} \sim \mathrm{N}\left(\mu, \left(\dfrac{\sigma}{\sqrt{n}}\right)^2\right)$
- Then, to convert $\bar{x}$ -value to z-score, the formula is: $z = \dfrac{\bar{x} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$
- If population SD, σ, is unknown, we can use sample SD, s, as an estimate for σ.
  But, then $\dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ follows a distribution known as **Student's *t*-distribution**.
  The use of *t*-distribution requires that the population is at least approximately normal.
  A feature of *t*-distribution is that it approaches the Z-distribution as *n* increases.
  We will elaborate on *t*-distribution in the next chapter.

Example 1: A lightbulb manufacturer claims that the lifespan of his lightbulbs follows a normal distribution with mean 750 hours and standard deviation 30 hours. A random sample of 20 lightbulbs is to be selected for testing of lifespan.

(a) Describe the sampling distribution of the sample mean lifespan of 20 lightbulbs.

(b) What is the probability that 20 lightbulbs will have a mean lifespan of 725 hours or less?

(c) Hence, is it rare to get a sample of 20 lightbulbs with a mean lifespan of 725 hours or less?

Example 2: Chocolate Delight produces chocolate bars for baking. The brand claims that their chocolate bars contain an average of 250g of cocoa content, with standard deviation of 20g. Amy, a baker, took a sample of 30 such chocolate bars and measured the cocoa content.

(a) Is the sampling distribution of the sample mean cocoa content of 30 chocolate bars normal? Why?

(b) What is the probability that the sample mean cocoa content differs from the claimed mean by at least 11g?

(c) Explain the meaning of this probability.

Case Study: Subway

Upon reading the news article which speculated that Subway's foot-long sandwiches are less than 12 inches, Thomas bought 30 foot-long Subway sandwiches and measured each of them. He recorded the readings as follows:

12.0	11.5	11.9	12.1	12.0	11.8	12.2	12.4	12.0	11.9
12.6	12.0	11.6	12.7	12.1	12.0	11.4	11.8	11.7	12.4
11.6	11.9	11.7	11.4	11.6	11.9	11.6	12.1	11.8	11.6

Thomas then used Minitab to calculate the summary statistics, given as follows:

Descriptive Statistics: Subway foot long sandwiches

Statistics

Variable	N	Mean	StDev
Subway foot long sandwiches	30	11.9100	0.32732

(a) What is the sample mean length and sample standard deviation of the 30 foot-long Subway sandwiches that Thomas bought?

(b) What is the claimed mean length of Subway's foot-long sandwiches? Take this claimed mean as the population mean.

(c) **Assume that the length of Subway's foot-long sandwiches is normally distributed.** What is the probability that the average length of 30 Subway's foot-long sandwiches measures at most the sample mean length that Thomas got?

(d) Interpret the probability calculated in part (c).

TUTORIAL 3A

1. Given that $\mu = 52$, $\sigma = 35$ and $n = 35$, what is $P(X > 73)$ and $P(\overline{X} > 73)$?

2. The paint drying time varies depending on the type of paint. The label on a can of a latex-based paint claims that the drying time is on average 200 minutes with a standard deviation of 50 minutes at room temperature. A random sample of 49 latex-based paint is examined for their drying times.
   (a) Describe the sampling distribution of the sample mean drying time of 49 latex-based paint.
   (b) Find the probability that the sample mean drying time is more than 210 minutes
   (c) Find the probability that the sample mean drying time is at most 160 minutes.
   (d) Find the probability that the sample mean drying time is between 180 and 220 minutes.

3. The following is taken from the website of DigitalTrend.com:

> According to an updated study released earlier today by Nielson, Facebook is still eats up the most amount of time for a typical Web user. Clocking in at 7 hours and 45 minutes, the average person spends that much time each month wandering though endless status updates, leaving comments on new media from friends and playing social games like Zynga's CityVille. However, Facebook is still number two when it comes to the total Internet audience. Google takes the top spot again and the average user spends about 1 hour and 45 minutes on Google products each month. AOL and its variety of Web properties take second place in regards to the amount of time spent on an AOL site, nearly three hours per user per month.
>
> *(http://tinyurl.com/digitaltrendsfb, retrieved on 2 January 2015)*

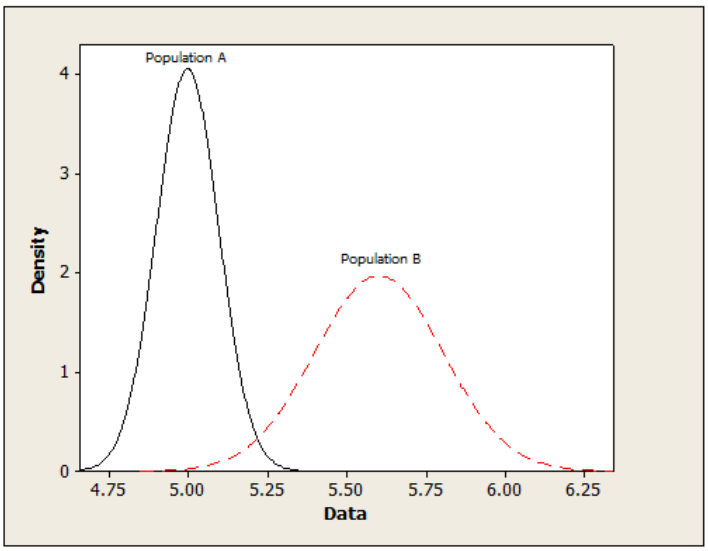
   (a) What is the average time spent (in minutes) by a Facebook user per month?
   (b) What is the probability that a Facebook user spends less than 300 minutes? Assume that time spent is normally distributed with $\sigma = 65$ minutes.
   (c) What is the probability that 40 students selected at random will spend an average of more than 500 minutes? Assume $\sigma = 65$ minutes.

4. A machine is regulated so that it produces a mechanical part with an average diameter of 240 mm and a standard deviation of 15 mm. This machine is routinely examined to ensure that it is working at the expected level. Periodically, a sample of 35 mechanical parts are checked and the mean diameter is computed. If the sample mean is within the interval of $\mu_{\overline{X}} \pm 2\,\sigma_{\overline{X}}$, then the machine is thought to be operating properly. Otherwise, adjustments will have to be made to the machine parts.
   (a) What is the range of values for $\mu_{\overline{X}} \pm 2\,\sigma_{\overline{X}}$?
   (b) In one of the regular checks, a quality control officer found that the mean of a sample of 35 mechanical parts to be 234 mm, and concluded that the machine needs adjustment. Using the answer from part (a), is his conclusion reasonable? Explain.

5. The weight of male students attending a large polytechnic is approximately normally distributed with $\mu = 68$ kg and $\sigma = 5$ kg. If twenty male students are crowded into the lift, what is the probability that the lift's maximum capacity of 1400 kg would be exceeded? Explain the meaning of the probability computed.

6. In a chemical process the amount of a certain type of impurity in the output is difficult to control. It is claimed that the population mean amount of the impurity is 0.2 g per gram of output. It is known that the standard deviation is 0.05 g per gram of output.
An experiment is conducted to gain more insight regarding the claim that $\mu = 0.2$ g.
The process was run on a lab scale 50 times and the sample average $\bar{x}$ turned out to be 0.23 g per gram of output. By computing an appropriate probability, comment on the claim that the mean amount of impurity is 0.2 g per gram of output.

7. You are given the following graphs:



$\mu = 4.998$

$\mu = 5.602$

$\sigma = 2.02$

$\sigma = 0.98$

(a) Two values of mean and two values of standard deviation are given but not labelled. Decide which values of mean and standard deviation best describe the distributions of Population A and Population B.

(b) A sample of 50 were taken from one of the populations. The sample mean is calculated to be 5.17. Determine whether this sample is more likely to be obtained from Population A or Population B. Justify your answer.

8. The manufacturer of ePads, an economy priced tablet, recently completed the design for a new tablet model. ePad's top management would like some assistance in pricing the new tablet. Two market research firms were contacted and asked to prepare a pricing strategy. Firm A tested the new tablet with 50 randomly selected consumers who indicated they plan to purchase a new tablet within the next year. Firm B test-marketed the new tablet with 200 current tablet owners. Which of the marketing research firm's test results will be more useful? Explain why.

ANSWERS

1. $P(X > 73)$ cannot be determined as the distribution of X is unknown.
   $P(\overline{X} > 73) = 0.0002$

2. (a) $\bar{X} \sim N\left(200, \left(\frac{50}{7}\right)^2\right)$ (b) 0.0808 (c) ≈0 (d) 0.9948

3. (a) 465 mins (b) 0.0055 (c) 0.0003

4. (a) 234.92 mm to 245.07 mm
   (b) Yes, because the sample mean of 234 mm is below the minimum amount of the acceptable range.

5. 0.0367. Since the probability is close to zero (<0.05), it is highly unlikely that we could get a sample mean of 20 male students from a population of weights with μ = 68 kg and σ = 5 kg that exceeds average the maximum lift capacity (in this case, it is 70kg).

6. $P(\overline{X} > 0.23) \approx 0$. Since the probability is close to zero (<0.05), it is rare that we could get a sample mean of 50 lab readings from a population of amount of impurities with μ = 0.2 g and σ = 0.05 g to be greater than 0.23 g. Hence, it is unlikely that we will believe the claim.

7. (a) Population A: μ = 4.998 (because the "midpoint" of the distribution is to the left relative to distribution of population B) and σ = 0.098 (because the spread is smaller compared to distribution of population B).
   (b) It is more likely to be from Population A because $P(Z > \frac{5.17-4.998}{0.98/\sqrt{50}}) = 0.1075$ compared to $P(Z < \frac{5.17-5.602}{2.02/\sqrt{50}}) = 0.0655$.

8. Larger samples provide narrower estimates of a population mean. So the company with 200 sampled customers can provide more precise estimates. In addition, they are selected consumers who are familiar with tablets and may be better able to evaluate the new tablet.

CHAPTER 3B
ESTIMATING POPULATION MEAN

Learning Objectives:

1. *Distinguish between population parameter and sample statistic.*
2. *Define point estimate.*
3. *Compute margin of error.*
4. *Construct confidence intervals from large samples with population standard deviation known or unknown.*
5. *Interpret confidence intervals.*
6. *Construct confidence intervals using Minitab by selecting Z or t-distributions.*

Content

1. Point Estimate and Confidence Interval

1.1 Estimating True Value

In statistical inference, we use **sample statistic** to estimate **population parameter**.

For examples:

Population parameter	Sample statistic
Patient's overall mean blood pressure	Nurse measures patient's blood pressure several times and take average
Average lifetime of a Duracell battery	Average lifetime of random Duracell batteries that are tested on the rack

The sample mean blood pressure is used to estimate the true value of the patient's mean blood pressure.

The sample mean lifetime of the tested Duracell batteries is used to estimate the true value of the lifetime of a Duracell battery on average.

But how well does the sample statistic really match the true value of the population parameter?

So, instead of using a single number (i.e. sample statistic), we can compute a range of values along with a confidence level for that range. This range of values is called a **confidence interval**.

Assumptions in constructing confidence intervals:

#1. _____ observations

#2. _____ distribution or sample size is sufficiently _____

#3. _____ population standard deviation σ

1.2 Point Estimate

Point estimate is a single number statistic that can be used to estimate a single number for the population parameter:

$\bar{x}$ = point estimate
μ = population parameter

This means that we use $\bar{x}$ as the point estimate for μ.

(Often, we use sample SD, s, as the point estimate for population SD, σ, when σ is unknown.)

One problem with using the sample mean $\bar{x}$ to infer the population mean μ is that $\bar{x}$ can vary depending on the sample we take. So, a single number like $\bar{x}$ is not a very helpful estimate of μ without some indication of how accurate it is.

1.3 Confidence Interval

Therefore, let's include a **margin of error**:

point estimate ± margin of error

Due to our assumptions, the sampling distribution of $\bar{x}$ is normal, then recall from the previous chapter that:

$$\mu_{\bar{X}} = \mu \ , \ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Applying the Empirical rule, then about 95% of sample means in the sampling distribution will lie within two standard deviations $\left(\sigma_{\bar{X}}\right)$ away from the mean $\left(\mu_{\bar{X}} = \mu\right)$.

However, μ is unknown, whereas $\bar{x}$ is known of the sample we took. So we say that we are "approximately 95% **confident** that μ is within $\bar{x} \pm 2\sigma_{\bar{X}}$".

That is how likely it is that the interval we come up with actually contains the unknown population mean μ. In other words, if we continue to take samples, we will catch the true value of μ about 95% of the time, over many samples.

2. Constructing Confidence Intervals

2.1 Known σ

In constructing confidence intervals, we can select a level of confidence that gives the probability that the estimation method will give an interval that catches the unknown population parameter (μ).

Confidence levels of 90%, 95% and 99% are usually chosen, with 95% being the most common.

Generally, confidence interval has the structure: **point estimate ± margin of error**

Specifically, for confidence interval of μ, the formula is: $\boxed{\bar{x} \pm z^*\left(\frac{\sigma}{\sqrt{n}}\right)}$

where **critical value *z**** takes the following values, obtained from z-table, and depending on selected confidence level:

Confidence level	90%	95%	99%
*z** value			

Case Study 1: Duracell Batteries

Reference: Confidence Intervals: Against All Odds—Inside Statistics. (2013). Films Media Group. Available at: http://fod.infobase.com/PortalPlaylists.aspx?wID=151497&xtid=111543

Battery companies, like Duracell, have always trumpeted their product's long lives in their commercials. Because the companies promise specific improvements in battery lifetimes, they need proof before the ads are aired.

At Kodack's Ultra Technologies, technicians use rigorous testing to back up the marketers' claims. Random samples of batteries are pulled from the warehouse and tested on the rack, which mimics the load of real products in a controlled environment.

Formulating questions	What is the average lifetime of Duracell's batteries?
Collecting data	_____ random Duracell batteries are tested on the rack. Let random variable X = _____ Sample mean lifetime, $\overline{x}$ = _____ mins Based on Kodack's past testing experiences, σ = _____ mins
Analysing data	μ = mean lifetime of Duracell's batteries Sampling distribution of $\overline{x}$ is normal because 95% confidence interval for μ is:
Interpreting results	

2.2 Unknown σ

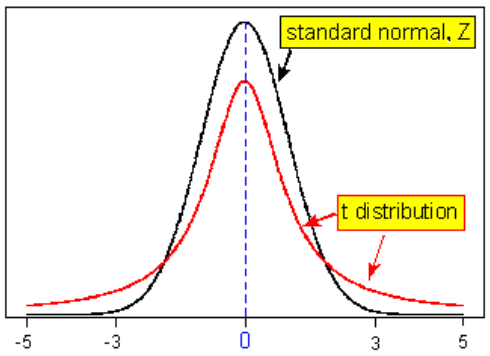
Usually, when the population mean μ is unknown, the population standard deviation σ is also unknown.

If the sample size is large enough, we can simply use sample standard deviation to stand in for population standard deviation.

However, if the sample size is small, we have to use the **Student's *t*-distribution** instead of Z-distribution for critical values. The use of *t*-distribution to compute confidence interval will be covered in practical.

Key features of the *t*-distribution:

- Centered at and symmetrical about 0
- More spread out than Z-curve
- Thicker tails than Z-curve
- Its shape depends on degree of freedom (d.f. = $n - 1$)
- As n increases, *t*-distribution approaches Z-distribution.



The decision to choose Z- or *t*-distribution for critical values is summarized in the table as follows:

Scenario	Critical value
If population X is normal and σ known.	Z*
If population X is normal and σ unknown.	t^* If n is large, $t^* \approx Z^*$
If population X is not normal, n is large and σ known.	Z*
If population X is not normal, n is large and σ unknown.	t^* Since n is large, $t^* \approx Z^*$

Case Study 2: Subway

Upon reading the news article which speculated that Subway's foot-long sandwiches are less than 12 inches, Thomas bought 30 foot-long subs and measured each of them. He recorded the readings as follows:

12.0	11.5	11.9	12.1	12.0	11.8	12.2	12.4	12.0	11.9
12.6	12.0	11.6	12.7	12.1	12.0	11.4	11.8	11.7	12.4
11.6	11.9	11.7	11.4	11.6	11.9	11.6	12.1	11.8	11.6

Assume that the length of Subway's foot-long subs is normally distributed.

Formulating questions

Is Subway's claim correct?

Collecting data

As mentioned, Thomas collected a sample of size $n = 30$.

Let random variable X = length of a foot-long sub.

Thomas used Minitab Express to compute the summary statistics as follows:

Descriptive Statistics: Subway foot long sandwiches

Statistics

Variable	N	Mean	StDev
Subway foot long sandwiches	30	11.9100	0.32732

Analysing data

μ = mean length of foot-long subs

Sampling distribution of $\overline{x}$ is normal because

95% confidence interval for μ is:

Interpreting results

TUTORIAL 3B

1. The American Management Association (AMA) wishes to have information on the mean income of middle managers in the retail industry. A random sample of 256 managers revealed a sample mean of $45,420. The standard deviation of this population is $2,050.
   (a) What is the population parameter? What is the sample statistic?
   (b) What is the point estimate for the population parameter?
   (c) Construct a 95% confidence interval for the population parameter.

2. Measurements of the diameters of a random sample of 200 ball bearings made by a certain machine showed a mean of 0.824 cm. Past data put population standard deviation of diameters at 0.042 cm.
   (a) Find the 90% confidence interval for the mean diameter of ball bearings. Interpret this interval.
   (b) If we construct confidence intervals by the same method 20 times, how many of these intervals will we expect to capture the true value of mean diameter?

3. Severe Acute Respiratory Syndrome (SARS) is a viral respiratory illness. It has the distinction of being the first new communicable disease of the 21st century. Researchers wanted to estimate the incubation period of patients with SARS. Based on interviews with 81 SARS patients, they found that the mean incubation period was 4.6 days with a standard deviation of 15.9 days. Using this information, construct a 95% confidence interval for the mean incubation period of the SARS virus.

4. The following summary gives the background characteristics of 50 participants in a research study evaluating HIV medications:

Patient characteristics	Mean (SD)
Age (years)	37.6 (6.8)
% male	75.1%
Education (years)	13.6 (2.4)
CD4 cell count (cells/ μl)	376 (94)

   Construct a 99% confidence interval for the mean CD4 cell count.

5. To encourage more shoppers in Orchard, the Urban Planning Authority (UPA) built a new multi-storey carpark that charges cheaply. UPA plans to pay for the structure through collected parking fees. During a two-month period (60 days), daily fees collected averaged $12,900, with a standard deviation of $1,650.
   (a) Construct a 95% confidence interval for the mean daily income this carpark will generate. Interpret this interval.
   (b) The consultant who advised UPA on this project predicted that parking revenues would average $13,500 per day. Based on your answer in part (a), do you think the consultant is correct? Explain.

6. In a factory, a sample of 50 resistors are randomly selected to estimate the true mean resistance of resistors produced by the factory. The 99% confidence interval for true mean resistance is computed to be between 98.6 Ω and 101.3 Ω.
   (a) What is the mean resistance of the sample?
   (b) What is the margin of error?
   (c) What is the standard error of the sample mean resistance?

7. A clinical trial was conducted to test the effectiveness of the drug zopiclone for treating insomnia in older subjects. Before treatment with zopiclone, 16 subjects had a mean wake time of 102.8 min. After treatment with zopiclone, the 16 subjects had a mean wake time of 98.9 min and a standard deviation of 42.3 min. Assume that the 16 sample values appear to be from a normally distributed population. Construct a 98% confidence interval estimate of the mean wake time for a population with zopiclone treatments. What does the result suggest about the mean wake time of 102.8 min before treatment? Does zopiclone appear to be effective?

ANSWERS

1. (a) Population parameter: mean income of middle managers in the retail industry
   Sample statistic: mean income of sample of 256 middle managers
   (b) $45,420 (c) $45,169 to $45,671
2. (a) I am 90% confident that the true value of the mean diameter of ball bearing is between 0.8191 cm and 0.8289 cm.
   (b) 18 times
3. 1.14 to 8.06 days, or 1.08 to 8.12 days
4. 341.8 to 410.2, or 340.4 to 411.6
5. (a) I am 95% confident that the mean daily income generated by the carpark is between $12,482.49 and $13,317.51 (or $12,473.76 to $13,326.24).
   (b) Not likely. $13,500 is not contained in the interval that I constructed.
6. (a) 99.95 Ω (b) 1.35 Ω (c) 0.524 Ω, or 0.504 Ω
7. 71.4 min to 126.4 min. The confidence interval includes the mean of 102.8 min that was measured before the treatment, so the mean could be the same after the treatment. This result suggests that the zopiclone treatment does not have a significant effect.

PRACTICAL 3B : Confidence Interval

Learning Objectives:

1. *Find confidence intervals using Minitab with raw data.*
2. *Find confidence intervals using Minitab with summarized statistics.*
3. *Select Z or t-distribution appropriately.*

Task 1

The boiling temperature (in °C) of a certain liquid is being studied. From past experiences, boiling temperature is known to be normally distributed with σ = 1.2 °C. A student measuring the boiling temperature on 6 different samples of the liquid observes the readings (in °C) to be 102.5, 101.7, 103.1, 100.9, 100.5 and 102.2.
Construct a 95% confidence interval for the mean boiling temperature.
Hence, interpret the confidence interval.

Task 2

Recall the following information from Case Study 1:

40 random Duracell batteries are tested on the rack.
Let random variable X = lifetime of a Duracell battery in minutes.
Sample mean lifetime, $\bar{x}$ = 450 mins
Based on Kodack's past testing experiences, σ = 63.5 mins.

Construct a 99% confidence interval for mean lifetime of Duracell batteries.
Hence, interpret the confidence interval.

Task 3

A machine is producing metal pieces that are cylindrical in shape. A sample of pieces is taken and the diameters (in cm) are:

1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, 1.03

Construct a 90% confidence interval of the mean diameter of pieces from this machine, assuming an approximate normal distribution.
Hence, interpret the confidence interval.

Task 4 (optional)

Find the critical values to construct the confidence intervals in tasks 1 to 3 above.

CHAPTER 4
HYPOTHESIS TESTING OF MEAN

Learning Objectives:

1. *Identify statistical claims.*
2. *Formulate or determine statistical hypotheses.*
3. *Describe hypothesis testing as an evaluation on the likeliness of obtaining the sample data assuming that the claim of a population parameter μ is true.*
4. *Conceptualize and interpret P-value through the informal concept of 'rare-events'.*
5. *Apply statistical problem solving, i.e., the systematic process of formulating questions, collecting data, analysing data and interpreting results.*
6. *Perform hypothesis testing of mean using Minitab.*
7. *Interpret Minitab outputs of hypothesis testing of mean and compare to confidence interval.*

Content

1. Case Study: Subway

In the previous chapter, we used the Subway case study to illustrate the concept of sampling distribution of the sample mean.

When you read the article about Subway foot-long sandwich measuring less than 12 inches, did you share one of the two responses below?

Maybe it is just BAD LUCK that the guy got this Subway sandwich.

Hmm… Could Subway be cheating its customers all these while…?

We will be examining both the responses in this chapter and why one response is better than the other. We will also perform a test, known as **Hypothesis Testing**, to check on Subway's claim using Thomas's sample data of 30 foot-long sandwiches with sample mean of 11.91 inches.

Let us first understand the logic underpinning the reasoning in a hypothesis test.
Most of the content in this chapter is adapted from the book entitled, *Stats Data and Model* by De Veaux, Velleman and Bock (2016).

2. Concept of Hypothesis Testing

The logic for hypothesis tests is similar to the logic of jury trials.

Let us suppose that a defendant has been accused of robbery and that the law states, everyone is innocent until proven guilty. By this logic, the defendant's status quo is innocent and the accuser has the burden of proof to show evidence that the status quo statement (e.g., the defendant is innocent) is not correct.

The aim is to reject the status quo statement with some kind of evidence, and offer an alternative statement (e.g., the defendant is guilty) to the status quo statement.

When the accuser finally presents the evidence to challenge the status quo (e.g., finding the defendant at the crime scene with a bag of money and wearing a mask at the time of the robbery), the next step is to judge the evidence.

The question to ask at this point is, if the status quo statement is assumed to be true, how *likely* would we observe this kind of evidence?

Using the jury trial analogy, let us go back to statistics.

Null Hypothesis	Ho: "status quo statement"
vs.	
Alternative Hypothesis	H1: "a statement to challenge status quo"

In a hypothesis test, the status quo statement is known as the _____ (denoted as $\mathbf{H_0}$) and the statement to challenge the null hypothesis is known as the _____ (denoted as $\mathbf{H_1}$ or $\mathbf{H_a}$).

H_0 is always assumed to be true unless someone wants to challenge it. H_0 can be challenged by offering a H_1 and evidence is needed. The evidence that we collect in a hypothesis test is the sample data.

Note that the hypotheses are statements about the population parameter.
In this course, the hypotheses made will always be about the population mean μ.

When sample data is presented, the next step is to evaluate the likelihood of getting such sample data under the assumption that H_0 is true. To do so, we use a measurement called the **P-value**.

In a nutshell, the hypothesis testing process resembles closely the statistical problem-solving process:

Step	Statistical Problem-Solving Process	**Hypothesis Testing**
1	Formulate questions	**Set up hypotheses**
2	Collect data	**Collect sample data**
3	Analyze data	**Compute P-value**
4	Interpret results	**Decide to reject H_0 or not**

3. Setting Up Hypotheses

Since null hypothesis should be a statement about status quo, it is usually written with the equality sign "=". For example,

$$H_0: \mu = 5$$

This means that under status quo, we know that the population mean is 5.

If this status quo is challenged, an alternative hypothesis has to be offered. In this case, there are three possibilities.

- If you think that the population mean is too high, then you offer that population mean should be smaller. That is,

  $$H_1: \mu < 5$$

  This will lead to a **one tail** (or **lower tail**) hypothesis test.

- If you think that the population mean is too low, then you offer that population mean should be higher. That is,

  $$H_1: \mu > 5$$

  This will also lead to a **one tail** (or **upper tail**) hypothesis test

- If you simply disbelieve that the population mean is 5, then you offer that population mean is not 5. That is,

  $$H_1: \mu \neq 5$$

  This will lead to a **two tail** hypothesis test

Example 1: In each of the contexts described below, set up the hypotheses.

Context	Hypotheses
(a) An adult needs at least 7 hours of sleep daily, but you do not think that Singaporeans get sufficient sleep.	*Status quo: they get sufficient sleep on average.* H_0: *Alternative: they do not get sufficient sleep on average.* H_1:
(b) The newspaper reported that the average income of Singaporean is \$2000, but you suspect the figure is off.	*Status quo: average income is \$2000.* H_0: *Alternative: average income is not \$2000.* H_1:
(c) A software company allows you to download a trial copy of their program and the instructions stated that the download will take 10 mins. However, you felt that it takes longer and you are frustrated by the wait.	*Status quo:* H_0: *Alternative:* H_1:

4. Assumptions

In order for hypothesis testing of mean to work, there are some assumptions that we have to make.

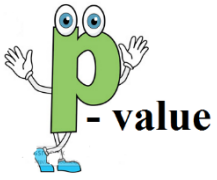
- Usually the standard deviation of the population in the study is not known. As such, we will have to use the *t*-distribution instead of the Z-distribution.
- But to use the *t*-distribution, the distribution of the population should still be at least approximately normal.
- If we are not sure about the population distribution, then we can rely on Central Limit Theorem to kick in if we take a sample size that is at least 30.

The decision to choose Z- or *t*-distribution for hypothesis testing is summarized in the table as follows (similar to *Chapter 4*):

If population X is normal and σ known.	Z-distribution
If population X is normal and σ unknown.	*t*-distribution If n is large, $t \approx Z$
If population X is not normal, n is large and σ known.	Z-distribution
If population X is not normal, n is large and σ unknown.	*t*-distribution Since n is large, $t \approx Z$

5. P-Value

The fundamental step in our reasoning is the question "Are our data surprising, given the null hypothesis?" The key calculation is to determine how likely the data we observed would be given that the null hypothesis is true.



Specifically, we want to find the *probability* of seeing data like these or something even less likely *given* that the null hypothesis is assumed true.

This probability tells us how SURPRISED we would be to see the data we have collected if the null hypothesis is true.

This probability is called the _____. It can be obtained easily from statistical software.

- When P-value is _____ it says that we are very surprised. It means that it is rare that we will observe sample data like these if the null hypothesis is true.

  The assumption that we have about the null hypothesis (which is H_0 is true) and the sample data that we have collected are in contradiction with each other. The assumption about null hypothesis and our sample data are not consistent.

  If we believe our sample data is collected and analyzed properly, then we have no choice but to reject our assumption about the null hypothesis. Formally, H_0 is _____.

- When P-value is _____, we have not seen anything that is unlikely or surprising at all if the null hypothesis is true. Events that have a high probability of happening happen often.

  The sample data collected are consistent with our assumption about the null hypothesis and hence we have no reason to reject the null hypothesis. Formally, H_0 is _____.

  However, note that we have not proven that our null hypothesis is surely true or correct. What a big P-value tells us is that the sample data that we collected did not provide evidence for us to doubt the assumption made on null hypothesis. The best thing we can say is, well, the null hypothesis does not appear to be wrong.

Case Study 2: Subway

A foot-long Subway sandwich is assumed to be, on average, 12 inches long (i.e. $\mu = 12$ inches).
From previous chapters, we know that Thomas collected a sample of 30 foot-long Subway sandwiches and found the sample mean, $\bar{x}$, to be 11.91 inches.
There seems to be evidence to challenge the status quo assumption about Subway's foot-long sandwich!

Step 1: Let us write down the hypotheses.

Status quo: on average, a foot-long Subway sandwich is 12 inches long
H_0:

Alternative: Thomas thinks Subway sandwich is not that long!
H_1:

Step 2: Collect sample data for evidence.

The length of a foot-long Subway sandwich is assumed to be normally distributed, as established in previous chapters.

Thomas have already collected $n =$ _____ Subway sandwiches.

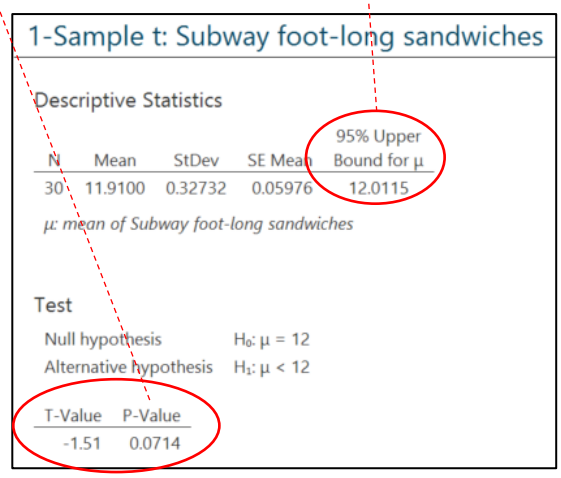
He found that the sample mean length, $\bar{x} =$ _____.

Step 3: Analyze the sample data by computing the P-value and confidence interval.

From *Chapter 3*, we have found that the probability of getting a sample mean length of at most ____ inches, assuming that population mean length is ___ inches, is approximately _____ (2 d.p.). This is actually the P-value!

When we enter the data into Minitab for analysis, the output shows that:

- P-value is _____
- The 95% confidence bound for μ is: $\mu \leq$ _____ inches.



Notice that the P-value is slightly different from the calculation that we did in the previous chapter. This discrepancy is due to the use of *t*-distribution instead of the Z-distribution as we did in the previous chapter. The *t*-distribution provides a more accurate calculation when the population standard deviation is unknown.

Step 4: Interpret the results and decide to reject status quo or not.

So, based on the P-value of 0.07, what can we say about the null hypothesis?

A P-value of 0.07 indicates that Thomas should not be all that surprised to a sample mean of 11.91 inches or less, assuming that the null hypothesis of $\mu = 12$ inches is true.

Furthermore, we are 95% confident that μ falls below 12.0115 inches, which includes the claimed $\mu = 12$.

There is insufficient evidence to doubt H_0, hence H_0 cannot be rejected based on Thomas's sample data.

Brief notes on confidence intervals and confidence bounds:

- When a two tail hypothesis test is performed, a symmetrical confidence interval is constructed. This confidence interval was covered in *Chapter 4*. We say that, at a confidence level, population mean μ is captured between a and b.
- When an upper tail hypothesis test is performed, a lower confidence bound is constructed. We say that, at a confidence level, population mean μ falls above a.
- When a lower tail hypothesis test is performed, an upper confidence bound is constructed. We say that, at a confidence level, population mean μ falls below b.

Coming back to these responses:

Maybe it is just BAD LUCK that the guy got this Subway sandwich.

Hmm… Could Subway be cheating its customers all these while…?

Comment:
If that is true, in statistics we would like to capture the probability of such an event happening. If bad luck happens too often, maybe there are other explanations.

Comment:
The best way to respond is like what Thomas did, he collected data in an attempt to challenge the status quo that the average Subway's foot-long sandwiches is l2 inches. But from his sample data, he realized that he does not have enough evidence to reject the status quo. Note that Thomas' data did not prove Subway is right (or honest) – his sample data just show that Subway does not appear to be cheating its customers (cannot reject H_0).

TUTORIAL 4

1. In each of the following article, do the following:
   - Identify the statistical claim made.
   - Identify the population and sample.
   - Formulate a set of hypotheses to be used in testing the stated claim.

(a)

S'pore youths spending more time online: study

By Fann Sim | Yahoo! Newsroom Fri, Jan 11, 2013

The average number of hours youths in the country spend online daily has gone up from 4.8 hours in 2011 to 5.5 hours last year, a study by Singapore Polytechnic found.

Results of the annual study, conducted in June last year, were based on person-administered surveys with 820 youths, defined as those aged between 15 and 35 years old.

(Retrieved from http://tinyurl.com/youthonline on 28 Jan 2015)

(b)

Average Singaporean works 2,287 hours a year: Study

AsiaOne
Monday, Sep 02, 2013

Singaporeans work some of the longest hours in the world's most developed countries, a study done by the Groningen Growth and Development Centre has shown.

According to statistics published on the Federal Reserve Economic Data (FRED) website, Singaporeans worked an average of 2,287 hours in 2011.

Globally, the average number of hours people worked ranged from 1,380 to 2,800 hours a year, with richer countries working relatively fewer hours.

(Retrieved from http://tinyurl.com/averagework on 28 Jan 2015)

(c)

Singapore ranks third globally in time spent on homework

15-year-olds here devote 9.4 hours weekly, above global average of 5 hours: OECD study

PUBLISHED ON DEC 25, 2014 9:22 AM
BY AMELIA TENG

Students in Singapore are among the world's most hard- working at home, clocking the third-longest time spent on homework, a report released this month has found.

The country's 15-year-olds said that they devoted 9.4 hours to homework a week, in the study by the Organisation for Economic Cooperation and Development (OECD).

They came in behind students in Shanghai, who spend 13.8 hours a week on homework, and those in Russia, who take 9.7 hours.

(Retrieved from http://tinyurl.com/straitstimeshw on 28 Jan 2015)

2. Question A: "Do non-smokers typically live longer than 70 years?"

   Question B: "Is there evidence to believe that Pizza Shed at Dover takes less than 30 minutes for delivery?"

   In each of the questions asked above, answer the following:

   - What is the population and sample?
   - What data will we collect?
   - Formulate a set of hypotheses to be used to answer questions A and B.

3. The packaging of a certain brand of potato chips stated a net weight of 28.5 g. Peter checked 6 packets and found their weight to be:

   29.2, 28.5, 28.7, 28.9, 29.1, 29.5

   He analyzed the data using Minitab and generated the following output:

1-Sample t: Weight

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
6	28.9833	0.3601	0.1470	(28.6054, 29.3612)

μ: mean of Weight

Test

Null hypothesis	H_0: μ = 28.5
Alternative hypothesis	H_1: μ ≠ 28.5

T-Value	P-Value
3.29	0.0218

   Is the stated net weight on the packaging correct?

   (a) What is the population and sample?

   (b) Set up the hypotheses to test if the stated net weight is correct.

   (c) Report and interpret the P-value and confidence interval.

   (d) What is the conclusion?

ANSWERS

1. (a) The average amount of time youth spend online daily is 5.5 hours (μ=5.5)
   Population: Singapore youths
   Sample: 820 Singapore youths aged between 15 to 35 years old
   H_0: $\mu = 5.5$, H_1: $\mu \neq 5.5$
   (b) The average yearly amount of hours worked by Singaporeans is 2287 (μ=2287)
   Population: Working Singaporeans
   Sample: Not mentioned
   H_0: $\mu = 2287$, H_1: $\mu \neq 2287$
   (c) On average, 15 year-olds spend 9.4 hours on homework ($\mu = 9.4$).
   Population: 15 year-old students in Singapore
   Sample: Not mentioned
   H_0: $\mu = 9.4$, H_1: $\mu \neq 9.4$

2. Question A
   Population – Non-smokers
   Sample – Non-smokers who take part in the study
   Data – No. of years lived/age at the time of death; quantitative data
   H_0: $\mu \leq 70$, H_1: $\mu > 70$, where μ is the average lifespan of non-smokers.

   Question B
   Population – Deliveries by Pizza Shed at Dover
   Sample – Deliveries in a particular month
   Data – Time taken for the delivery; quantitative data
   H_0: $\mu \geq 30$ mins , H_1: $\mu < 30$ mins , where μ is the average delivery time taken.

3. (a) Population – all packets of potato chips of a certain brand
   Sample – 6 packets of potato chips of a certain brand checked by Peter
   (b) H_0: $\mu = 28.5$g , H_1: $\mu \neq 28.5$g , where μ is the mean weight of a packet of potato chips
   (c) P-value is 0.0218. It is rare to obtained the sample mean weight as far as 28.98g if $\mu = 28.5$g. Furthermore, 95% CI for μ is: $28.6\text{g} < \mu < 29.4\text{g}$. The claimed net weight of 28.5g falls outside of the confidence interval.
   (d) There is evidence to reject H_0 and conclude that the stated net weight of the packaging is not correct.

PRACTICAL 4 : Hypothesis Testing

Learning Objectives:

1. *Perform hypothesis testing using Minitab with raw data.*
2. *Perform hypothesis testing using Minitab with summarized data.*

Task 1:

<Data set can be found in "STAT_Prac4_Data" Excel worksheet (compiled from http://support.minitab.com/en-us/datasets/).>

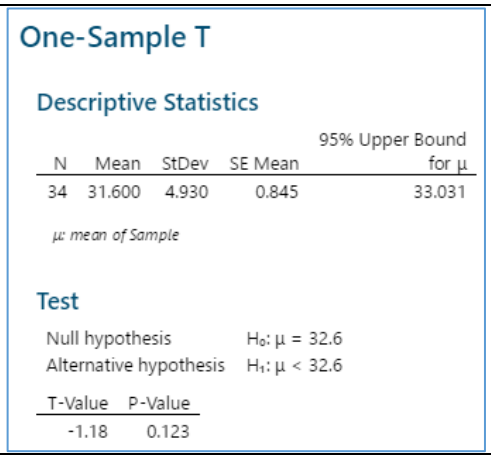
A packaging engineer needs to ensure that the force required to open snack bags is within the target value of 4.2 N. The engineer tests the force required to open 28 bags where the data were recorded as the force required to open the snack bags (in Newton). On average, does the force required to open the snack bags meet the target value? Assume that force required is normally distributed.

1. Formulating questions	On average, _____ Let X = _____, and μ = _____ H_0 : H_1 :
2. Collecting data	Sample size, n = Sample mean, $\bar{x}$ = Sample SD, s =
3. Analyzing data	Since X is _____ and σ is _____, use _____. Use Minitab to analyse the data: P-value = _____ 95% confidence interval for μ is: _____ 1-Sample t: Force Descriptive Statistics N Mean StDev SE Mean 95% CI for μ 28 4.4850 0.7319 0.1383 (4.2012, 4.7688) μ: mean of Force Test Null hypothesis H_0: μ = 4.2 Alternative hypothesis H_1: μ ≠ 4.2 T-Value P-Value 2.06 0.0491
4. Interpreting results	The P-value obtained is _____, which is _____. This means that it is ____ to get a sample mean force of _____, if the engineer had assumed that the population mean force, μ, is _____. Furthermore, the engineer is 95% confident that the mean force is captured in _____, which _____ the assumed population mean of _____. Hence, there is _____ evidence to reject H_0 based on the sample obtained by the engineer. On average, _____ _____.

Task 2

A governing body regulates taxis' fuel economy and sets an annual fuel mileage of 32.6 miles per gallon (mpg) or better for taxis. To continue to meet the regulation, Tuber Cabs checked the petrol usage for 34 taxis selected randomly from their fleet, finding a mean of 31.6 mpg and a standard deviation of 4.93 mpg. Is there evidence that Tuber Cabs has ceased to meet regulation? Assume that mileage is normally distributed.

1. Formulating questions	_____? Let X = _____, and μ = _____. H_0 : H_1 :
2. Collecting data	Sample size, n = Sample mean, $\bar{x}$ = Sample SD, s =
3. Analyzing data	Since X is _____ and σ is _____, use _____. Use Minitab to analyse the data: P-value = _____ 95% upper confidence bound for μ is: _____.
4. Interpreting results	P-value = _____ is _____. This means that it is _____ to get a sample mean mileage of _____ _____, if Tuber Cabs' mean mileage is assumed to be _____. Furthermore, the 95% upper confidence bound for mean mileage is _____ _____, which includes the assumed population mean of _____. Hence, there is _____ evidence to reject H_0 based on the sample obtained by Tuber. _____.



One-Sample T

Descriptive Statistics

N	Mean	StDev	SE Mean	95% Upper Bound for μ
34	31.600	4.930	0.845	33.031

μ: mean of Sample

Test

Null hypothesis H_0: μ = 32.6
Alternative hypothesis H_1: μ < 32.6

T-Value	P-Value
-1.18	0.123

CHAPTER 5
CONCEPTS OF HYPOTHESIS TESTING

Learning Objectives:

1. *Define significance level.*
2. *Identify critical value(s) corresponding to a significance level and the type of test.*
3. *Compute test statistic and P-value.*
4. *Understand the relationship between test statistic and P-value.*
5. *Interpret from both critical value method and P-value method.*
6. *Recognise the risk of making Type I and Type II errors in hypothesis testing.*

Content

1. Significance Level & Critical Values

In the previous chapter, we learnt what P-value is. The P-value tells us the probability of getting results at least as unusual as the observed statistic, given the null hypothesis is true.

When the P-value is smaller than 0.05, it means that it is very rare that we will observe such sample data if the null hypothesis is true. In this scenario, we reject the null hypothesis.

When the P-value is larger than 0.05, it means that we are not surprised to observe the sample data if the null hypothesis is true. In this case, we do not reject the null hypothesis.

The 0.05 level is commonly used as the threshold for rare events. This threshold is called the **significance level** or **alpha (α) level**. Other common significance levels are 0.10 and 0.01.

Significance level, α, is a probability and it is also the complement of confidence level.
That is,

$$\alpha = 1 - \text{confidence level}$$

In chapter 4, we learnt that a confidence interval is a range of values that is likely to contain an unknown population parameter. If we draw random samples many times and construct similar confidence intervals, a certain percentage of the confidence intervals will include the population parameter. This percentage is the confidence level.

Significance level, α		5 %	
Confidence level, 1 – α			

The significance level is used to determine certain lower and/or upper limits of the normal curve. These limits are called **critical values** and they determine the **rejection regions**, also known as **critical regions**, in a hypothesis test. These critical values are the same ones used to construct confidence intervals in chapter 3B. Assuming normal distribution and population standard deviation is known, we use the Z-distribution and the critical value is denoted by z^*.

Taking two tail test as illustration:

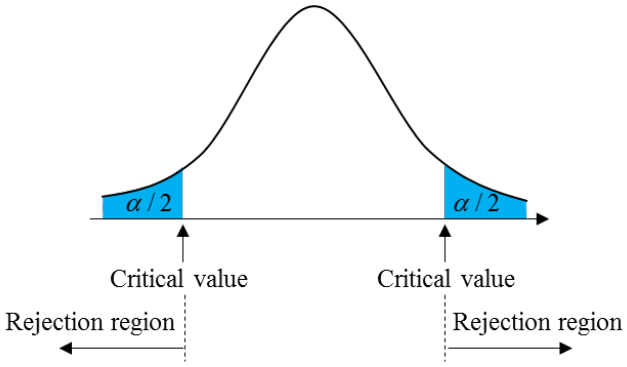
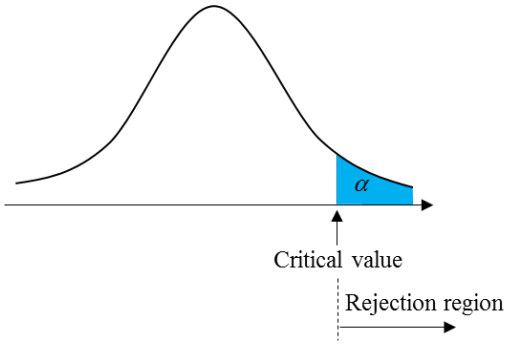
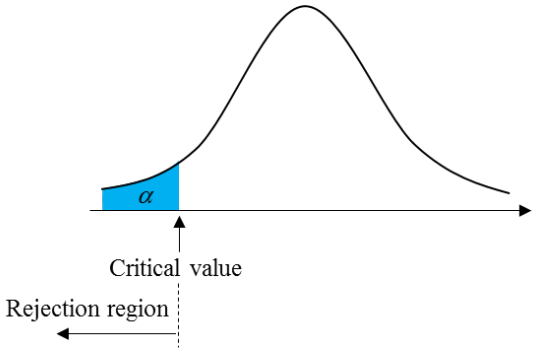
A 95% confidence interval can be written as,

$$\mathrm{P}\left(\bar{X} - z^* \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z^* \frac{\sigma}{\sqrt{n}} \right) = 0.95$$

$$\mathrm{P}\left(-z^* \frac{\sigma}{\sqrt{n}} < \mu - \bar{X} < z^* \frac{\sigma}{\sqrt{n}} \right) = 0.95 \quad \Rightarrow \mathrm{P}\left(-z^* \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z^* \frac{\sigma}{\sqrt{n}} \right) = 0.95$$

$$\mathrm{P}\left(-z^* < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < z^* \right) = 0.95$$

Any claimed μ_0 that can be captured within the interval is considered plausible.

A <u>two tail test</u> has both lower and upper limits. The critical values split α equally into two tails, such that the area of each tail is $\frac{\alpha}{2}$. These critical values are used in constructing the confidence intervals, as well as determining the rejection regions in a hypothesis test.	
An <u>upper tail test</u> has upper limit but no lower limit. The critical value puts all α on the right side.	
A <u>lower tail test</u> has lower limit but no upper limit. The critical value puts all α on the left side.	

Example 1: Using the Z-distribution, fill in the critical values in the table below, corresponding to the significance levels and type of tests.

	Type of test		
significance level, α	lower tail	two tail	upper tail
0.01			
0.05			
0.10			

2. Test Statistic

From the sample obtained, we can compute sample statistic such as sample mean. We can then compute the z-score for the sample mean, which is known as the **test statistic** using:

$$z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

where $\overline{x}$ is the sample mean

μ_0 is the assumed population mean in the null hypothesis (hypothesized mean)

σ is the population standard deviation

n is the sample size

If the test statistic is more extreme than the critical value(s) or falls in the rejection region, the null hypothesis will be rejected. This is an alternative to using the P-value for hypothesis testing.

Recall in Chapter 5 where we learnt the following steps involved in the hypothesis testing:

Step	Statistical Problem-Solving Process	Hypothesis Testing
1	Formulate questions	Set up hypotheses
2	Collect data	Collect sample data
3	Analyse data	Compute P-value
4	Interpret results	Decide to reject H_0 or not

Case Study 1: Weight of Teenagers

A study conducted 10 years ago found the population mean weight and standard deviation of teenagers to be 70kg and 10kg respectively. A researcher believes that the mean weight of 70 kg is no longer valid since the study was conducted 10 years ago. He feels that the mean weight has changed but he is not sure if it has increased or decreased. He obtained the weights of a sample of 36 teenagers and found their mean to be 75kg. Perform a hypothesis test at significance level of 0.05 to see if the researcher's belief is correct.

Step 1: Write down the hypotheses.

Step 2: Collect sample data for evidence.

The researcher obtained weights from a sample of

The have a mean weight

Population standard deviation was

Step 3: Analyse the sample data.

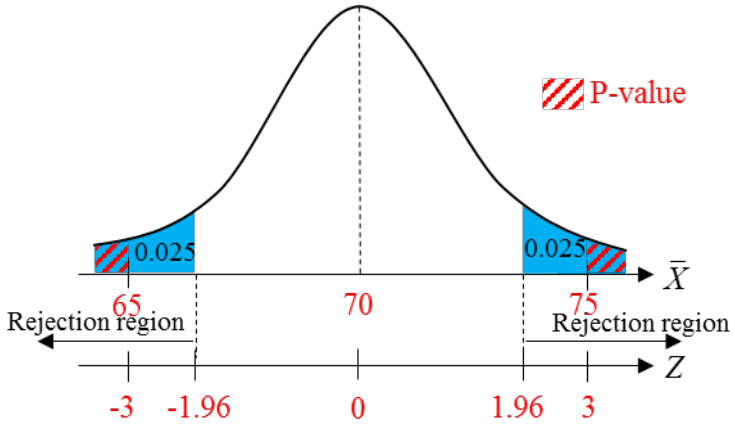
By Critical Value Method	By P-value Method
Critical value, z* =	Manual computation of P-value = 2 × P(Z > 3)
Test statistic, z =	

Output from Minitab:

One-Sample Z

Test of μ = 70 vs ≠ 70
The assumed standard deviation = 10

N	Mean	SE Mean	95% CI	Z	P
36	75.00	1.67	(71.73, 78.27)	3.00	0.003



Step 4: Interpret the results and decide to reject status quo or not.

By Critical Value Method	By P-value Method
Since test statistic = 3 is more extreme than the critical values, it is rare to obtain a sample mean weight of 75 kg or more, if the population mean weight is 70 kg.	Since P-value = 0.003 < α = 0.05, it is rare to obtain a sample mean weight of 75 kg or more, if the population mean weight is 70 kg.

Hence, H_0 is rejected at α = 5%. The mean weight of teenagers has changed.

Usually, the population standard deviation σ is unknown. In this case, we have to use the sample standard deviation *s* to estimate σ. Consequently, we have to use the Student's *t*-distribution to determine the test statistic:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \text{ with } n - 1 \text{ degrees of freedom (df)}$$

When *t*-distribution is used, the critical values is denoted by *t**. If the test statistic is more extreme than the critical value(s) or falls in the rejection region, the null hypothesis will be rejected.

3. Errors

Nobody's perfect. Even with lots of evidence, we can still make the wrong decision. When we perform a hypothesis test, we can make mistakes in two ways:

- The null hypothesis is true, but we mistakenly reject it.
- The null hypothesis false, but we fail to reject it.

These two types of errors are known as **Type I error** and **Type II error** respectively. They can be summarised by the following matrix:

	H_0 is true	H_0 is false
Do not reject H_0		
Reject H_0		

How often will a Type I error occur? It happens when the null hypothesis is true, but we had unfortunately drawn an unusual sample. To reject H_0, the P-value must fall below α. When H_0 is true, that happens *exactly* with probability α. So when we choose the level α, we are setting the probability of a Type I error to α.

In Case Study 1 (since H_0 is rejected), a Type I error would occur if the researcher concluded that the mean weight of teenagers has changed, when actually the mean population weight of teenagers is still 70 kg.

Case Study 2: Contamination in Farmed Salmon

In 2004, a team of researchers published a study on contaminants in farmed salmon. Fishes from many sources were analysed for 14 organic contaminants. The study expressed concerns about the level of contaminants found. One of those was the insecticide mirex, which has been shown to be carcinogenic and is suspected to be toxic to the liver, kidneys and endocrine system.

The researchers tested 150 farm-raised salmon for organic contaminants. They found the mean concentration of mirex to be 0.0913 parts per million (ppm), with standard deviation 0.0495 ppm. The authority recommended that mirex concentration should not exceed 0.08 ppm. At 0.05 significance level, are the farmed salmon contaminated beyond the level permitted by the authorities?

Step 1: Write down the hypotheses.

Let X = _____, and μ = _____

H_0 :

H_1 :

Step 2: Collect sample data for evidence.

The researchers collected a sample of _____ farmed-raised salmons.

Mean level of mirex contamination was

Standard deviation of the sample was

Step 3: Analyse the data sample.

By Critical Value Method	By P-value Method

Step 4: Interpret the results and decide to reject status quo or not.

By Critical Value Method	By P-value Method
Since test statistic = ____ is ___ the critical region, it is _____ to obtain a sample mean mirex concentration of _____ _____, if the population mean mirex concentration is _____.	Since P-value = _____ $\alpha = 0.05$, it is ____ to obtain a sample mean mirex concentration of _____, if the population mean mirex concentration is _____.

Hence, H_0 is _____ at $\alpha = 5\%$. _____ _____.

Type I or Type II error?

A _____ error would have occurred if the mean concentration of mirex is actually _____ the contamination level permitted, but the researchers concluded _____.

Case Study 3: Mean Debt-to-Equity Ratio

One measure of a company's financial health is its debt-to equity ratio. This quantity is defined to be the ratio of the company's corporate debt to the company's equity. If this ratio is too high, it is one of the indications of financial instability. Banks often monitor the financial health of companies to which they have extended commercial loans. In order to reduce risk, a bank has decided to initiate a policy limiting the mean debt-to-equity ratio for its portfolio of commercial loans to being less than 1.5.

The bank randomly selects a sample of 15 of its commercial loan accounts. These 15 accounts have a mean debt-to-equity ratio of 1.343 with standard deviation of 0.192. At 0.05 significance level, can the bank ascertain that the mean debt-to-equity ratio of its commercial loans is less than 1.5? Assume that debt-to-equity ratio is normally distributed.

Step 1: Write down the hypotheses.

Let X = _____, and μ = _____

H_0 :

H_1 :

Step 2: Collect sample data for evidence.

Sample size:

Sample mean debt-to-equity ratio:

Sample SD of debt-to-equity ratio:

Step 3: Analyse the sample data.

By Critical Value Method	By P-value Method

Step 4: Interpret the results and decide to reject status quo or not.

By Critical Value Method	By P-value Method

It is _____ to obtain a sample mean debt-to-equity ratio of _____, if the population mean debt-to-equity ratio is _____.

Hence, H_0 is _____ at $\alpha = 5\%$. _____ _____.

Type I or II error?

Since H_0 is _____, there is a __ chance of committing _____ error. A _____ error would have occurred if the mean debt-to-equity ratio is _____, but the bank concluded _____.

TUTORIAL 5

1. For each of the following, find the critical value(s) z^*:

   (a) $H_0 : \mu = 10$ vs $H_1 : \mu \neq 10$ at $\alpha = 0.05$.

   (b) $H_0 : \mu = 20$ vs $H_1 : \mu < 20$ at $\alpha = 0.01$.

   (c) $H_0 : \mu = 106$ vs $H_1 : \mu > 106$ at $\alpha = 0.10$.

2. Find the t critical value(s) for $\alpha = 0.05$ and $n = 15$ when the alternative hypothesis is

   (a) $H_1 : \mu \neq 0$.

   (b) $H_1 : \mu > 0$

   (c) $H_1 : \mu < 0$

3. A Statistics lecturer has observed that for several years students scored an average of 105 points out of 150 on the semester exam. A salesman suggests that he try a statistics software package that gets students more involved with computers, predicting that it will increase students' scores. The software is expensive, and the salesman offers to let the lecturer use it for a semester to see if the scores on the final exam increase significantly. The lecturer will have to pay for the software only if he chooses to continue using it. In that semester, 203 students signed up for the statistics course. They use the software suggested by the salesman, and scored an average of 108 points with a standard deviation of 8.7 points. At 0.05 significance level, should the lecturer spend money on the software?

   (a) State the null and alternative hypotheses.

   (b) Identify the critical value.

   (c) Compute the test statistic and state your conclusion.

4. In a survey, 583 workers were asked how many hours they worked in a week. The mean was 37 hours with a standard deviation of 15.1 hours. At 0.01 significance level, does this suggest that the mean number of hours worked is significantly different from 40 hours?

   (a) State the null and alternative hypotheses

   (b) Identify the critical value.

   (c) Compute the test statistic and state your conclusion.

5. An industrial plant claims to discharge no more than 1000 gallons of wastewater per hour, on the average, into a neighbouring lake. An environmental action group decides to monitor the plant in case this limit is being exceeded. Doing so is expensive, and only a small sample is possible. A random sample of four hours is selected over a period of a week. The mean amount of wastewater discharged is 2000 gallons with standard deviation of 816.5 gallons. Assume that amount of wastewater discharged is normally distributed.

   (a) Write the null and alternative hypotheses.

   (b) Identify the critical value. Assume 0.10 significance level.

   (c) Compute the test statistic and state your conclusion.

   (d) Find the P-value.

6. In a recent study, a group of moderately obese subjects were randomly assigned one of the three diets: low-fat, restricted-calorie; Mediterranean, restricted-calorie; or low-carbohydrate, non-restricted calorie. It is suspected that subjects taking low-carbohydrate, non-restricted calorie diet would show changes in weight. The MINITAB output below shows results of a hypothesis test:

One-Sample T

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
109	-5.500	7.000	0.670	(-6.829, -4.171)

μ: mean of Sample

Test

Null hypothesis	H_0: μ = 0
Alternative hypothesis	H_1: μ ≠ 0

T-Value	P-Value
-8.20	0.000

   (a) State the null and alternative hypotheses.

   (b) State the significance level.

   (c) Identify the P-value and state your conclusion.

7. The housing market crashed in an economic crisis in 2008. A census taken in 2010 showed a mean loss of $10200. In 2012, realtors randomly sampled 31 bids from potential buyers to estimate the average loss in home value. The sample showed the average loss from the peak in 2008 was $9560 with a standard deviation of $1500. At 0.05 significance level, has the mean loss decreased?

   (a) State the null and alternative hypotheses.

   (b) Identify the critical value.

   (c) Compute the test statistic and state your conclusion.

8. A tyre manufacturer is considering a newly designed tread pattern for its all-weather tyres. Tests have indicated that the tyres will provide better petrol mileage and longer tread life. The last test for the tyres is testing for its braking effectiveness. The company hopes that the tyre will allow a car travelling at 100 km/h to come to a complete stop within an average of 40 m. They will adopt the new tread pattern unless there is strong evidence that the tyres do not meet this objective. From a sample of 10 stops on a test track, the mean braking distance was 40.5 m with standard deviation of 2.8 m. Take 0.05 as the significance level and assume that braking distance is normally distributed.

   (a) State the null and alternative hypotheses.

   (b) Identify the critical value.

   (c) Compute the test statistic and state your conclusion.

   (d) Find the P-value.

9. An insurance company is reviewing its current policy rates. When the rates were originally set, they believed the average claim amount was $1800. The policy was not well-received and they are now concerned that they might have overstated the true claim amount, thus resulting in uncompetitive premiums. They randomly selected 40 claims and found the mean amount to be $1650 with standard deviation of $500. They carried out a hypothesis test and obtained the following output in MINITAB:

   (a) State the null and alternative hypotheses.

   (b) State the significance level.

   (c) Show that the test statistic is -1.90 and state your conclusion.

One-Sample T

Descriptive Statistics

N	Mean	StDev	SE Mean	90% Upper Bound for μ
40	1650.0	500.0	79.1	1753.1

μ: mean of Sample

Test

Null hypothesis	$H_0: \mu = 1800$
Alternative hypothesis	$H_1: \mu < 1800$

T-Value	P-Value
-1.90	0.033

10. To justify the construction of a wind turbine to generate electricity, a study found that the site should have a mean annual wind speed above 13 km/h. One candidate site was monitored for a year. A hypothesis test was carried out and the following output was obtained in MINITAB:

   (a) State the null and alternative hypotheses.

   (b) What is the sample size, sample mean wind speed and sample standard deviation?

   (c) State the P-value and conclusion.

One-Sample T

Descriptive Statistics

N	Mean	StDev	SE Mean	95% Lower Bound for μ
1114	13.210	6.200	0.186	12.904

μ: mean of Sample

Test

Null hypothesis	$H_0: \mu = 13$
Alternative hypothesis	$H_1: \mu > 13$

T-Value	P-Value
1.13	0.129

11. For each of the following situations, state whether a Type I, Type II, or neither error has been made.

    (a) A bank wants to know if the mean account balance is more than \$25,000. They take a sample of 200 account balances to test $H_0 : \mu = \$25,000$ vs. $H_1 : \mu > \$25,000$ and reject the null hypothesis. Later they find out that the mean balance in all accounts is actually \$24,900.

    (b) A student took the height of 100 students to determine if their average height is 1.7 m and finds no evidence that their average height is different from 1.7 m. Later the school measures the height of all students and finds no difference.

    (c) A human resource analyst wants to know if the applicants this year score, on average, higher on their placement exam than 52.5 points the candidates averaged last year. She samples 50 recent tests and finds the average to be 54.1 points. She fails to reject the null hypothesis that the mean is 52.5 points. At the end of the year, they find that the candidates this year had a mean of 55.3 points.

ANSWERS

1. (a) $z^* = \pm 1.96$ (b) $z^* = -2.33$ (c) $z^* = 1.28$
2. (a) ±2.145 (b) 1.761 (c) -1.761
3. (a) $H_0 : \mu = 105$, $H_1 : \mu > 105$ (b) $t^* = 1.652$ (c) 4.91
4. (a) $H_0 : \mu = 40$, $H_1 : \mu \neq 40$ (b) $t^* = \pm 2.584$ (c) -4.80
5. (a) $H_0 : \mu = 1000$, $H_1 : \mu > 1000$ (b) $t^* = 1.638$ (c) 2.45 (d) 0.046
6. (a) $H_0 : \mu = 0$, $H_1 : \mu \neq 0$ (b) 0.05 (c) 0
7. (a) $H_0 : \mu = \$10200$, $H_1 : \mu < \$10200$ (b) $t^* = -1.697$ (c) -2.38
8. (a) $H_0 : \mu = 40$ m, $H_1 : \mu > 40$ m (b) $t^* = 1.833$ (c) 0.56 (d) 0.293
9. (a) $H_0 : \mu = \$1800$, $H_1 : \mu < \$1800$ (b) 0.10
10. (a) $H_0 : \mu = 13$ km/h, $H_1 : \mu > 13$ km/h
    (b) $n = 1114$, $\bar{x} = 13.21$ km/h, $s = 6.2$ km/h (c) 0.129
11. (a) Type I (b) No error (c) Type II

PRACTICAL 5 : Concept of Hypothesis Testing

Minitab Practice:

A school technology committee has the opinion that the average time spent by students per lab visit has increased, and the increase supports the need for increased lab fees. To substantiate this opinion, the committee randomly samples 12 student lab visits and notes the amount of time spent using the computer. The time in minutes are as follows:

52	**57**	**54**	**76**	**62**	**52**
74	**53**	**80**	**73**	**50**	**62**

The previous mean amount of time spent using the lab computer was 55 minutes. At 0.05 significance level, what do you conclude about the claim? Assume amount of time spent is normally distributed.

<table>
<tr><td>1. Formulating questions</td><td colspan="2">Let X =
and μ =
H_0 :
H_1 :</td></tr>
<tr><td>2. Collecting data</td><td colspan="2">Using Minitab to display descriptive statistics,
• Sample size:
• Sample mean amount of time:
• Sample SD of amount of time:
n _____, σ _____, use _____</td></tr>
<tr><td>3. Analyzing data</td><td><u>Critical Value method:</u></td><td><u>P-Value method:</u></td></tr>
<tr><td rowspan="2">4. Interpreting results</td><td></td><td></td></tr>
<tr><td colspan="2">It is ____ to obtain a sample mean amount of time of _____, if the population mean amount of time is _____.
Hence, H_0 is _____ at α = 5%. _____ _____.
Type I or II error?
Since H_0 is _____, there is a ___ chance of committing _____ error. A _____ error would have occurred if the average time spent has _____ _____, but the committee concluded _____.</td></tr>
</table>

CHAPTER 6
HYPOTHESIS TESTING OF TWO MEANS

Learning Objectives:

1. *Distinguish between independent and dependent samples when comparing two means.*
2. *Formulate null and alternative hypotheses.*
3. *Select the appropriate hypothesis tests for independent samples versus dependent samples.*
4. *Perform paired t-test using statistical software.*
5. *Perform two-sample independent t-tests using statistical software.*
6. *Interpret results to make statistical inference.*

Content

1. Independent Samples vs. Dependent Samples

In practice, it is common to compare two populations or two treatments from which the samples are drawn. For instance, the following scenarios can be analyzed by comparing two samples:

- Do brand A tablets have longer battery life than brand B tablets?
- Is a new diet programme effective in managing weight?
- Is a new drug effective in reducing high blood pressure?
- Are online book prices lower than retail prices at a local bookstore?

Notice that in the scenarios above, some samples are taken from two different populations, while some samples are taken from the same population, *twice*. Thus, it is important to distinguish between **independent** and **dependent** samples before formulating any hypotheses.

Definition:

- Two samples are **independent** if the values from one sample are not related to or paired to the values from another sample.
- Two samples are **dependent** if each value from one sample is paired to or matched with another value from another sample.
  Dependent samples are also called **paired** samples.

Example 1: For each of the following scenarios, state the samples collected and determine whether they are independent or dependent.

(a) A consumer compares the battery life of brand A and brand B tablets.	Sample 1: Sample 2: Independent / Dependent ?
(b) A dietitian enrolls 30 volunteers in a new diet program. All volunteers will take the same diet for a week and their weights are measured before and after the program.	Sample 1: Sample 2: Independent / Dependent ?
(c) A researcher randomly assigns 30 patients with hypertension to either control group or treatment group. Patients in the treatment group take a new drug everyday while those in control group take a placebo. The blood pressure of patients in both groups are measured and compared after one month.	Sample 1: Sample 2: Independent / Dependent ?
(d) Alice compares the online prices and retail prices of 30 textbooks.	Sample 1: Sample 2: Independent / Dependent ?

2. Paired *t*-test

Once two samples are determined to be dependent or paired, a new variable *d*, that is the difference between the values of each pair, can be defined as:

$$d = X - Y$$

where *d* is normally-distributed.

Usually the data can be tabulated in columns as follows:

Observation	X	Y	$d = X - Y$
1	x_1	y_1	$d_1 = x_1 - y_1$
2	x_2	y_2	$d_2 = x_2 - y_2$
⋮	⋮	⋮	⋮
n	x_n	y_n	$d_n = x_n - y_n$
Mean	$\overline{x}$	$\overline{y}$	$\overline{d}$

If there is no difference between the paired samples, then we would expect that on average, the differences will be zero. A **paired *t*-test** can then be used to determine whether the **mean of the differences** (μ_d) between paired samples differs significantly from zero.

The null and alternate hypotheses can be written as follows:

Two tail test	Lower tail test	Upper tail test
H_0:	H_0:	H_0:
H_1:	H_1:	H_1:

We will never be able to know the population standard deviation of all possible differences (σ_d). Hence, we always use sample standard deviation of differences (s_d) instead. As such, *t*-test is always used for comparing paired samples. Furthermore, the sample size (*n*) is the number of pairs, not the total number of data.

We can also construct a confidence interval to estimate the mean difference (μ_d).
For example, if a 95% confidence interval includes 0, then we are 95% confident that the interval captures 0. Hence, we do not reject null hypothesis at 5% significance level.

Example 2: Suppose that a new medicine is claimed to reduce blood sugar level among diabetes patients. A group of researchers measures the blood sugar level of 30 diabetes patients before and after being given the medicine. Formulate a set of hypotheses to test this claim.

Let X_{before} be the blood sugar level of patient before taking the medicine.
Let X_{after} be the blood sugar level of patient after taking the medicine.

Let d =

H_0: H_1:

Case Study 1: Weight Management Program

A new chain of fitness centre, Republic Fitness (RF), patented a weight management program which combines exercise and diet plans. RF claims that the 4-weeks program is effective in helping their members reduce weight. To verify the claim, RF randomly selected 30 participants to go through the program for free.

Step 1: Write down the hypotheses.

Samples are dependent (paired) because

Let difference d =

And let μ_d =

Since RF is interested to test whether the new weight management program is effective in reducing weight, the difference in weight before and after the program is expected to be positive/negative on average.

Status quo: program is not effective in reducing weight

H_0:

Alternative: program is effective in reducing weight

H_1:

Step 2: Collect sample data for evidence.

The weights of 30 participants before and after completing the program are recorded as follows:

Participant	Weight before (kg)	Weight after (kg)
1	65.8	63.2
2	59.3	60.6
3	78.4	70.1
…	…	…
30	80.7	78.9

[*The full data set can be downloaded from eSP Practical folder*]

So, n = _____

Step 3: Analyze the sample data using Minitab.

Checking for normality: since we do not know if the populations are normally distributed, we check if the samples might have come from a normal distribution (*check this!*).

Perform a _____ tail, paired *t*-test (d.f. = ____) at $\alpha = 0.05$.

(*Can you find the critical region?*)

Minitab output of paired *t*-test:

Paired T-Test and CI: Weight_Before, Weight_After

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Weight_Before	30	62.88	10.50	1.92
Weight_After	30	61.79	10.55	1.93

Estimation for Paired Difference

Mean	StDev	SE Mean	95% Lower Bound for μ_difference
1.090	3.208	0.586	0.095

μ_difference: mean of (Weight_Before - Weight_After)

Test

Null hypothesis H_0: μ_difference = 0
Alternative hypothesis H_1: μ_difference > 0

T-Value	P-Value
1.86	0.036

The output shows that the difference in weights, '*Difference'*, will be computed automatically in Minitab based on the values of '*Weight Before*' and '*Weight After*'.

$\bar{d} =$ $s_d =$

test statistic = P-value =

The 95% confidence bound for μ_d is:

<u>Step 4</u>: Interpret the results and decide to reject status quo or not.

P-value = 0.036 < $\alpha = 0.05$, indicates that it is <u>rare</u> to get a sample mean difference in weight of at least 1.09 kg if null hypothesis $\mu_d = 0$ kg is true.

(*Furthermore, test statistic = 1.86 is in the rejection region.*)

Also, the confidence bound <u>does not include</u> the hypothesized mean difference of 0 kg.

So, we <u>reject</u> H_0 at $\alpha = 5\%$.

Hence, RF can conclude that their weight management program is effective in significantly reducing weight on average.

3. Independent *t*-test

Suppose that samples are independently selected from two normal-distributed populations. If the population standard deviations are known, we can use hypothesis test based on the Z-distribution. However, since the population standard deviations are usually unknown (as with the population means), we will use **independent two-sample *t*-test** to make inference about the difference between the means of two populations.

Let μ_1 be the mean of first population, and μ_2 be the mean of second population.
If there is no difference between the two independent populations, then we would expect that their populations means will be equal $(\mu_1 = \mu_2)$. In other words, the difference in their population means $(\mu_1 - \mu_2)$ will be zero.

The null and alternate hypotheses can be written as follows:

Two tail test	Lower tail test	Upper tail test
H_0: H_1:	H_0: H_1:	H_0: H_1:

Note that the sample sizes drawn from the two populations are not necessarily the same. However, vastly unequal sample sizes could possibly lead to unequal variances.

Example 3: A business analyst is interested to find out if there is any difference in online shopping expenditure between male and female customers in the month leading up to Valentine's Day. He randomly surveyed 30 male customers and 32 female customers on their online shopping spending in that month. Formulate a suitable hypothesis test for this study.

Let μ_{male} be the mean online spending of male customers in that month.
Let μ_{female} be the mean online spending of female customers in that month.

H_0: H_1:

There are two sub-types of independent two-sample *t*-test:

- Two-sample *t*-test with **equal variances**.
  Since the population standard deviations or variances are equal, the standard deviations or variances will be **pooled**. The degree of freedom for the *t*-distribution is $n_1 + n_2 - 2$.
- Two-sample *t*-test with **unequal variances**.
  In this case, the *t*-distribution and its degree of freedom are both approximated.

A hypothesis test (*F*-test) can be conducted to determine if there is significant difference in two population variances. However, this test is beyond the scope of this module. In this chapter, problem statement will indicate if equal variances can be assumed.

Case Study 2: Blood Pressure

At a research clinic, a patient with systolic blood pressure reading between 140 and 159 mmHg will be classified as having Stage 1 Hypertension (S1H). A researcher wants to show that a new drug is effective in lowering the blood pressure of S1H patients. He randomly assigned 15 S1H patients to *treatment* group and another 15 S1H patients to *control* group. (By randomly assigning patients to treatment group and control group, the researcher removed biases in selection.)
Patients in the treatment group were administered the new drug, whereas patients in the control group were given a placebo. After one month, the blood pressure of these patients are measured and compared.

Step 1: Write down the hypotheses.

Samples are independent because

Let μ_1 =

And let μ_2 =

Since researcher wants to test whether the new drug lowers blood pressure, the mean of blood pressure of treatment group is expected to be higher/lower than the mean of blood pressure of control group.

Status quo: new drug is not effective in lowering blood pressure

H_0:

Alternative: new drug is effective in lowering blood pressure

H_1:

Step 2: Collect sample data for evidence.

The blood pressure of 15 S1H patients in treatment group and 15 S1H patients in control group is recorded as follows:

Blood Pressure of Control Group	Blood Pressure of Treatment Group
142	135
144	140
136	128
…	…
132	136

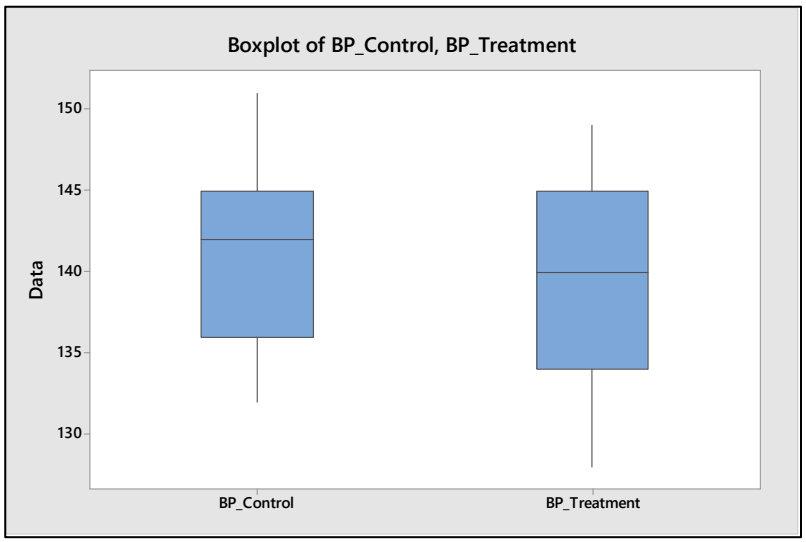
[*The full data set can be downloaded from eSP Practical folder*]

So, n_1 = _____ , n_2 = _____

<u>Step 3</u>: Analyze the sample data using Minitab.

Minitab output of descriptive statistics (numerical & graphical summaries):

Variable	Total Count	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
BP_Control	15	140.87	6.06	132.00	136.00	142.00	145.00	151.00	9.00
BP_Treatment	15	139.13	6.35	128.00	134.00	140.00	145.00	149.00	11.00



Numerical summaries show that the mean, median, minimum and maximum blood pressure of treatment group are all slightly lower than that of control group. Both boxplots are fairly symmetrical without any outliers. There is much overlap between these two boxplots. The SD, range and IQR of treatment group are slightly larger than those of control group.

Checking for normality: since we do not know if the populations are normally distributed, we check if the samples might have come from a normal distribution (*check this!*).

Perform a _____ tail, independent *t*-test at $\alpha = 0.05$.
Since the standard deviations of both groups are quite similar, ***assume* that the variances are equal.** Hence, select two-sample *t*-test with equal variances.
So, d.f. = _____

(*Can you find the critical region?*)

Minitab output of two-sample *t*-test with equal variances:

Two-Sample T-Test and CI: BP_Control, BP_Treatment

Method

μ_1: mean of BP_Control
μ_2: mean of BP_Treatment
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
BP_Control	15	140.87	6.06	1.6
BP_Treatment	15	139.13	6.35	1.6

Estimation for Difference

Difference	Pooled StDev	95% Lower Bound for Difference
1.73	6.20	-2.12

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
0.77	28	0.225

$\bar{x}_1 =$ $\bar{x}_2 =$ $\bar{x}_1 - \bar{x}_2 =$ $s_{pooled} =$

test statistic = P-value =

The 95% confidence bound for $\mu_1 - \mu_2$ is:

<u>Step 4</u>: Interpret the results and decide to reject status quo or not.

P-value = 0.225 > α = 0.05, indicates that it is <u>not rare</u> to get a difference of sample mean blood pressures of at least 1.73 mmHg if null hypothesis $\mu_1 - \mu_2 = 0$ mmHg is true.

(*Furthermore, test statistic = 0.77 is in the acceptance region.*)

Also, the confidence bound <u>includes</u> the hypothesized difference of means of 0 mmHg.

So, we <u>do not reject</u> H_0 at α = 5%.

Hence, the researcher failed to conclude that the new drug is effective in significantly lowering blood pressure on average.

Case Study 3: Flipped Classroom

In a traditional (usually higher-education) classroom, students will attend lectures in school and complete hands-on practices at home. In a "flipped classroom", students will watch lectures via videos at home and complete hands-on practices in class instead. Proponents of this recently-popular pedagogy feel that more customized attention and guidance can be given to students in class, especially needed when they are stuck when attempting practices.

In order to test whether this new pedagogy is effective, Miss Tan implemented flipped classroom for one class (Class 01), while another class (Class 02) went through traditional mode of learning.

At the end of a semester, Miss Tan compared the exam marks of these two classes. Assume unequal variances.

Step 1: Write down the hypotheses.

Samples are independent because

Let μ_1 =

And let μ_2 =

Status quo: flipped classroom is not more effective than traditional mode

H_0:

Alternative: flipped classroom is more effective than traditional mode

H_1:

Step 2: Collect sample data for evidence.

There are 20 students in Class 01 and 19 students in Class 02. At the end of the semester, the exam marks of the students are recorded as follows:

Class 01	Class 02
87	75
65	89
76	71
…	…

[*The full data set can be downloaded from eSP Practical folder*]

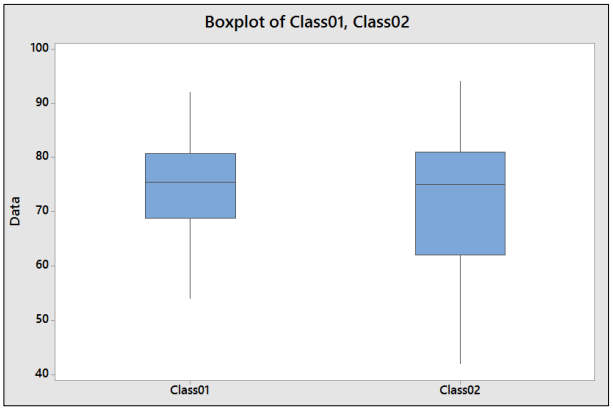
So, n_1 = _____ , n_2 = _____

Step 3: Analyze the sample data using Minitab.

Minitab output of descriptive statistics (numerical & graphical summaries):

Statistics

Variable	Total Count	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Class01	20	74.95	9.10	54.00	68.75	75.50	80.75	92.00	12.00
Class02	19	71.84	15.18	42.00	62.00	75.00	81.00	94.00	19.00



Both boxplots are quite symmetrical with no outlier. The mean exam marks of Class01 is higher than that of Class02, but the medians are almost the same. Class01 also has smaller variation (IQR, range, SD) as compared to Class02. Most importantly, the standard deviations of these two classes are quite different, suggesting that the population variances are not equal.

Checking for normality: since we do not know if the populations are normally distributed, we check if the samples might have come from a normal distribution (*check this!*).

Two-Sample T-Test and CI: Class01, Class02

Method

μ_1: mean of Class01
μ_2: mean of Class02
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Class01	20	74.95	9.10	2.0
Class02	19	71.8	15.2	3.5

Estimation for Difference

Difference	95% Lower Bound for Difference
3.11	-3.75

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
0.77	29	0.224

$\bar{x}_1 =$ $\bar{x}_2 =$ $\bar{x}_1 - \bar{x}_2 =$ d.f. $\approx$

test statistic = P-value =

The 95% confidence bound for $\mu_1 - \mu_2$ is:

Step 4: Interpret the results and decide to reject status quo or not.

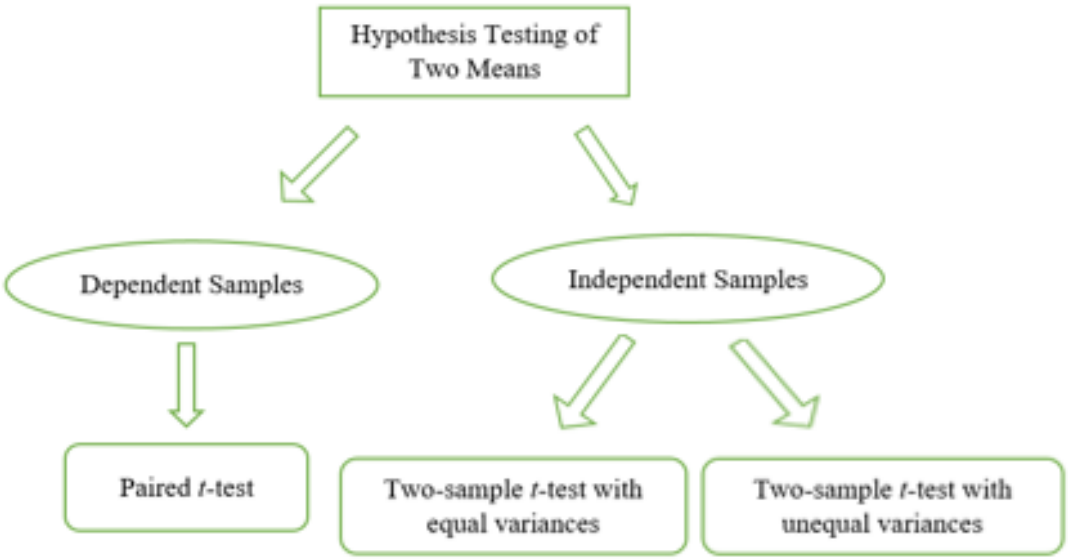
P-value = 0.224 > α = 0.05, indicates that

Also, the confidence bound

So, we _____ H_0 at α = 5%.

Hence, Miss Tan

4. Summary



TUTORIAL 6

1. Researchers are interested to find out whether on average, husbands are older than their wives. Data are collected from a random sample of 100 married couples. Assume that the ages of husbands and wives are normally distributed.
   (a) Give an example of how you would tabulate the data collected. (Show the first 3 rows of data.)
   (b) Are the samples collected dependent or independent? Explain.
   (c) Set up the null and alternative hypotheses, and select a suitable hypothesis test.
   (d) Suppose that the conclusion is "there is sufficient evidence to reject the null hypothesis at $\alpha = 0.05$". State a possible P-value and a 95% confidence interval.
   (e) Suppose that the conclusion is "there is insufficient evidence to reject the null hypothesis at $\alpha = 0.01$". State a possible P-value and a 99% confidence interval.

2. A team of business analysts is interested to find out if there is any significant difference in consumer spending on a trip to NTUC FairPrice or Sheng Siong in the same HDB estate. Assume that consumer spending is normally distributed.
   (a) Suggest a way to collect data so that the two samples are independent.
   (b) Suppose that a 95% confidence interval for the mean difference between NTUC FairPrice and Sheng Siong customer spending is calculated to be $(-\$4.95, -\$1.23)$. Explain in context what this interval mean.
   (c) Assuming the variances are equal, state the null and alternate hypotheses, and select a suitable hypothesis test.
   (d) Based on the hypotheses in part (c), suppose that the P-value is calculated to be 0.03, interpret the P-value and state the conclusion. Is this conclusion consistent with that of the 95% confidence interval in part (b)?

3. A team of investment analysts compares the annualized rates of return (%) of two investment portfolios over a period of 12 quarters. Portfolio A consists of common stocks only while Portfolio B consists of government bonds only. Assume normality.
   (a) Based on Figure 1 below, which portfolio shows higher volatility? Give two reasons to justify your answer.

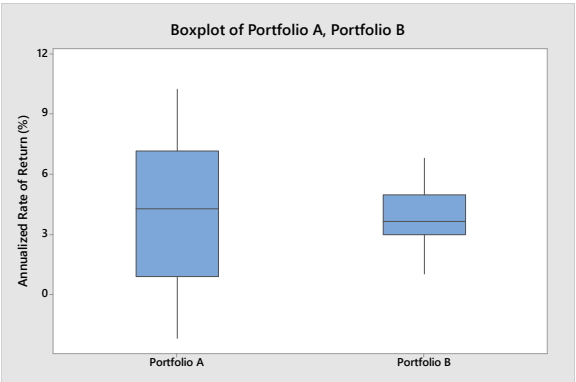


Figure 1: Boxplots of Portfolio A and Portfolio B.

(b) The team of investment analysts consists of Amy, Becky and Chris. Each of them formulates different set of hypotheses to test the claim that "common stocks will give higher average return on investment".
Who formulated the correct hypotheses based on their Minitab outputs below?
Justify your answer and explain the mistakes made by other analysts.

Two-Sample T-Test and CI: Portfolio A, Portfolio B

Method

μ_1: mean of Portfolio A
μ_2: mean of Portfolio B
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Portfolio A	12	3.93	3.88	1.1
Portfolio B	12	3.90	1.55	0.45

Estimation for Difference

Difference	Pooled StDev	95% CI for Difference
0.03	2.95	(-2.47, 2.53)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
0.02	22	0.980

Figure 2: Amy's analysis

Two-Sample T-Test and CI: Portfolio A, Portfolio B

Method

μ_1: mean of Portfolio A
μ_2: mean of Portfolio B
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Portfolio A	12	3.93	3.88	1.1
Portfolio B	12	3.90	1.55	0.45

Estimation for Difference

Difference	95% Lower Bound for Difference
0.03	-2.09

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
0.02	14	0.490

Figure 3: Becky's analysis

Paired T-Test and CI: Portfolio A, Portfolio B

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Portfolio A	12	3.93	3.88	1.12
Portfolio B	12	3.90	1.55	0.45

Estimation for Paired Difference

Mean	StDev	SE Mean	95% Lower Bound for μ_difference
0.03	4.71	1.36	-2.41

μ_difference: mean of (Portfolio A - Portfolio B)

Test

Null hypothesis $H_0: \mu\_difference = 0$
Alternative hypothesis $H_1: \mu\_difference > 0$

T-Value	P-Value
0.02	0.492

Figure 4: Chris's analysis

(c) Based on the correct hypotheses above, what can the team of investment analysts conclude about the claim?

4. A company organizes a free exercise programme for its employees to determine if it will improve their job satisfaction, as measured by a survey. The survey scores for randomly selected employees before and after the implementation of the exercise programme are analyzed and summarized in the Minitab output below.

Assume that survey scores are normally distributed.

(a) How many employees participated in this survey?

(b) State the mean survey scores before and after the exercise programme.

(c) State the type of hypothesis test performed by the company.

(d) Based on the Minitab output above, state the null and alternative hypotheses.

(e) Interpret the P-value and confidence bound. What is the company's conclusion?

Paired T-Test and CI: Before, After

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Before	10	27.00	8.67	2.74
After	10	35.50	9.10	2.88

Estimation for Paired Difference

Mean	StDev	SE Mean	95% Upper Bound for μ_difference
-8.50	7.47	2.36	-4.17

μ_difference: mean of (Before - After)

Test

Null hypothesis H_0: μ_difference = 0

Alternative hypothesis H_1: μ_difference < 0

T-Value	P-Value
-3.60	0.003

5. Researchers wanted to determine whether carpeted and uncarpeted rooms contained the same amount of bacteria. To determine the amount of bacteria in a room, researchers pumped the air from each room over a Petri dish for 8 carpeted and 8 uncarpeted rooms. Colonies of bacteria were allowed to form in the 16 Petri dishes over one week. The result of the experiment is summarized in the Minitab output below. Assume normality.

(a) State the mean bacteria count in carpeted and in uncarpeted rooms.

(b) State the type of hypothesis test performed by the researchers. Assume equal variances.

(c) Based on the Minitab output above, state the null and alternative hypotheses.

(d) State and interpret the P-value and confidence interval. What is the researchers' conclusion?

Two-Sample T-Test and CI: Carpeted, Uncarpeted

Method

μ_1: mean of Carpeted
μ_2: mean of Uncarpeted
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Carpeted	8	11.20	2.68	0.95
Uncarpeted	8	9.79	3.21	1.1

Estimation for Difference

Difference	Pooled StDev	95% CI for Difference
1.41	2.96	(-1.76, 4.58)

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
0.95	14	0.356

ANSWERS

1. (a)

Couple	Age of husband	Age of wife	Difference in age
1	32	30	2
2	34	35	-1
3	40	35	5
…	…	…	…

   (b) The samples are dependent as each couple is paired up.
   (c) Paired *t*-test. $H_0: \mu_d \le 0$ vs. $H_1: \mu_d > 0$
   (d) If the null hypothesis is rejected at $\alpha = 0.05$, then a possible P-value is 0.03.
   A possible confidence bound is $\mu_d > 1$.
   (e) If the null hypothesis failed to be rejected at $\alpha = 0.01$, then a possible P-value is 0.03.
   A possible confidence bound is $\mu_d > -1$.
   (*Remark: Other valid answers are accepted.*)

2. (a) The team of analyst can collect spending data from randomly chosen *m* NTUC FairPrice customers and *n* Sheng Siong customers by asking their purchase amount as they leave the stores. All customers sampled in both shops are different.
   (b) Since the 95% confidence interval does not include 0, we are 95% confident that the average customer spending are different for both shops.
   (c) Two tail two-sample *t*-test assuming equal variances.
   $H_0: \mu_{NTUC} - \mu_{SS} = 0$ vs. $H_1: \mu_{NTUC} - \mu_{SS} \neq 0$
   (d) Since P-value < $\alpha = 0.05$, H_0 is rejected.

3. (a) The boxplot of Portfolio A shows higher range and IQR than that of Portfolio B. Hence, Portfolio A has higher volatility (variation) as compared to Portfolio B.
   (b) Amy's analysis is wrong because she should not use two tail test to test this claim. and she should not assume equal variances. Chris is wrong because he should not use paired *t*-test. Becky is right to choose a one tail, two-sample *t*-test, without assuming equal variances. $H_0: \mu_A - \mu_B \le 0$ vs. $H_1: \mu_A - \mu_B > 0$
   (c) P-value > α =0.05, do not reject H_0.

4. (a) 10
   (b) 27, 35.5
   (c) Lower tail paired *t*-test.
   (d) $H_0: \mu_d \ge 0$ vs. $H_1: \mu_d < 0$
   (e) Since P-value = 0.003 < $\alpha = 0.05$, reject H_0.

5. (a) 11.20, 9.79
   (b) Two-sample *t*-test, assumed equal variance, two tail test.
   (c) $H_0: \mu_{carpeted} - \mu_{uncarpeted} = 0$ vs. $H_1: \mu_{carpeted} - \mu_{uncarpeted} \neq 0$
   (d) P-value = 0.355, 95% confidence interval is $(-1.76, 4.58)$ bacteria per cubic foot.
   Since the P-value > $\alpha = 0.05$ and the confidence interval includes 0, H_0 is not rejected.

Practical 6 : Hypothesis Testing of Two Means

[Data can be downloaded from eSP *Practical* folder.]

Task 1

A researcher wishes to show that a new drug is effective in lowering the blood pressure of stage 1 hypertension patients (systolic blood pressure 140 – 159). Instead of assigning patients to treatment group and control group, the researcher treats all 30 patients with the new drug. Two measurements of blood pressure are taken from each patient, one before the treatment and another 1 month after treatment with the new drug.

(a) Use Minitab to generate numerical and graphical summaries.
   i. State the means and standard deviations of blood pressure of these patients before and after treatment.
   ii. Describe the shape of distribution of blood pressure of these patients before and after treatment.

(b) Are these two samples dependent or independent? Explain

(c) State the type of hypothesis test which is suitable for this problem.

(d) Formulate the null and alternative hypotheses.

(e) Use Minitab to perform the hypothesis test to check normality and interpret the result.

Task 2

A recent study investigates if there are any differences in body mass index (BMI) values between males and females who are considered high risk for coronary heart disease.

(a) Use Minitab to generate numerical and graphical summaries.
   i. How many males and females participated in this study?
   ii. State the means and standard deviations of BMI of males and females.
   iii. Based on a suitable graph, describe the shape of distribution of BMI of males and females.

(b) Are these two samples dependent or independent? Explain.

(c) State the type of hypothesis test which is suitable for this problem.

(d) Formulate the null and alternative hypotheses.

(e) Use Minitab to perform the hypothesis test to check normality and interpret the result.

Task 3

An environmental impact study investigates if there are any differences in the mean noise level of various jets at a new airport. The noise level in decibel (dB) of narrow-bodied and wide-bodied jets were measured immediately after takeoff.

(a) Use Minitab to generate numerical and graphical summaries.
   i. State the means and standard deviations noise level of narrow-bodied jets and wide-bodied jets in this study.
   ii. Describe the shape of distribution of noise level of narrow-bodied jets and wide-bodied jets in this study.

(b) Are these two samples dependent or independent? Explain.

(c) State the type of hypothesis test which is suitable for this problem.

(d) Formulate the null and alternative hypotheses.

(e) Use Minitab to perform the hypothesis test to check normality and interpret the result.

Task 4

As a price conscious consumer, you would like to find out if the prices of business textbooks at local bookstore are different from the prices offered by online retailer Amazon.
Test an appropriate hypothesis and state your conclusion.

Answers

If your hypotheses are correct, you should get the following P-values:

1. Paired *t*-test, upper tail test, P-value < 0.001.
2. Two-sample *t*-test, assumed equal variance, two tail test, P-value = 0.023.
3. Two-sample *t*-test, do not assume equal variance, two tail test, P-value = 0.005.
4. Paired *t*-test, two tail test, P-value = 0.033.

CHAPTER 7
ANALYSIS OF VARIANCE (ANOVA)

Learning Objectives:

1. *Understand the concept behind a one-way ANOVA.*
2. *Perform one-way ANOVA using statistical software.*
3. *Interpret the results of a one-way ANOVA.*

Content

References
Moore David S., McCabe George, P, and Craig Bruce A (2009): *Introduction to the Practice of Statistics*, Freeman and Company NY.
Sullivan M. (2007): *Statistics: Informed Decisions Using Data*. Second Edition. Pearson Education.

1. Introduction to ANOVA

Previously, we had used two-sample *t*-tests to compare two population means. We now extend the concept of comparing two population means to comparing two or more population means. The procedure for doing this is called **Analysis of Variance**, or **ANOVA** for short.

In an ANOVA, the **response** variable is usually quantitative and the intent of the ANOVA is to determine the effect of one or more **factor** variables on the response. The factor can be quantitative or qualitative, and the **levels** of a factor refer to the possible values which the factor can take on.

The response is also known as outcome variable or dependent variable.
The factor is also known as explanatory variable, grouping variable or independent variable.

In this chapter, we will focus on **one-way** ANOVA. The term *one-way* indicates that there is a single factor, with two or more levels, employed in the analysis. The levels of the factor are also referred to as the **treatments** when a single factor is employed in an ANOVA.

Example 1: Suppose we are interested in investigating whether students from different courses perform differently in Mathematics. Groups, consisting of 6 students each, were randomly selected from four different courses – Design, Business, Engineering and Life Sciences. The 24 students took the same Mathematics paper and their scores were recorded. *Do students in different courses perform differently in Mathematics*?

What is the response?

What is the factor?

How many levels?

What are the levels?

Example 2: (*Refer to video posted in eSP on 'MoneyGuess'.*)
The psychological experiment is conducted to find out if holding different weighted clipboards affect people's estimation of money in a jar.

What is the response?

What is the response data type?

What is the factor?

What is the factor data type?

How many levels?

What are the levels?

2. Assumptions

Certain assumptions have to be met in order to conduct a one-way ANOVA.
These assumptions, simplified, are:

1. Each sample is an independent random sample drawn from different populations.
2. The distribution of the response variable in each population is normal.
3. Equal variances (or standard deviations) among all populations.

Statistical software has functions that allow the above assumptions to be checked. For example, in Minitab, we can produce residual plots to verify independence, normality and (constant) variance.

If one or more of the assumptions are violated, the conclusion drawn from the results may not be valid. In such cases, we can transform the data or perform a nonparametric test instead.

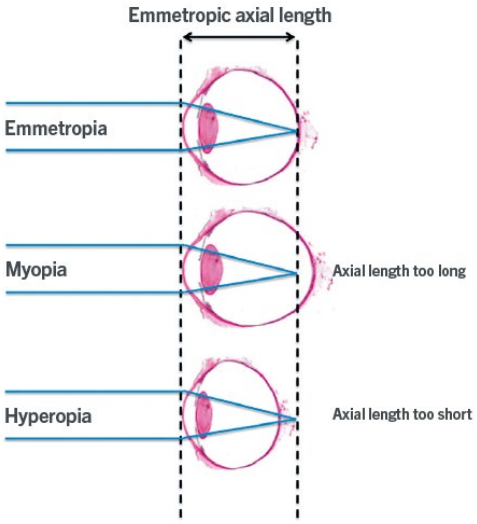
Informal rule-of-thumb to check assumption #3:

If $\frac{\text{largest sample SD}}{\text{smallest sample SD}} < 2$, then we assume that the population standard deviations are approximately equal.

Case Study: Myopia

[Download data set from eSP *Practical* folder]

References:
1. Wong TY, Loon SC, Saw SM (2006). The epidemiology of age related eyes diseases in Asia.
2. Wong TY, Foster PJ, Johnson GJ, Seah SK (2003). Refractive errors, axial ocular dimensions, and age related cataract: the Tanjong Pagar Survey.



Wong, *Loon* and *Saw* in their article *The epidemiology of age related eye diseases in Asia* reported a high prevalence of myopia in children and young adults in many Asian countries including Singapore when compared with children in many Western countries. A trend of declining prevalence of myopia with age was also observed. Furthermore, the Tanjong Pagar Survey showed that the difference in myopia rates between younger and older people could largely be explained by longer axial length of the eye in younger people.

A group of students from the Diploma in Optometry programme conducted a project to investigate the association between axial length and myopia of the eye. A sample of 64 volunteers aged 17 to 55 years was recruited. Axial length of the eye was recorded using the IOLMaster (Carl Zeiss Meditec). The results obtained are given in the following table:

Group	*Axial Length (mm)*							
High Myopia	26.22	25.72	25.91	27.70	26.26	27.32	27.44	28.28
	25.68	26.06	26.66	26.46	25.95	25.41		
Moderate Myopia	24.78	23.88	25.13	25.65	24.94	22.09	25.37	24.66
	24.06	24.88	24.22	23.09	23.23	25.19	24.14	25.05
	24.75	24.57	24.45	24.53	24.54	24.41	25.94	25.38
	25.33	23.62	24.59	25.85				
Normal	24.59	23.46	23.18	23.16	23.45	22.44	23.04	23.97
	22.95	23.20	23.81	23.50	23.47	23.44	23.20	23.50
	23.81	22.55	23.09	23.32	23.82	24.83		

Does the mean axial length of the eye differ in different myopia groups in the population?

Example 3: Refer to *Case Study*.

(a) State the response, factor and levels.

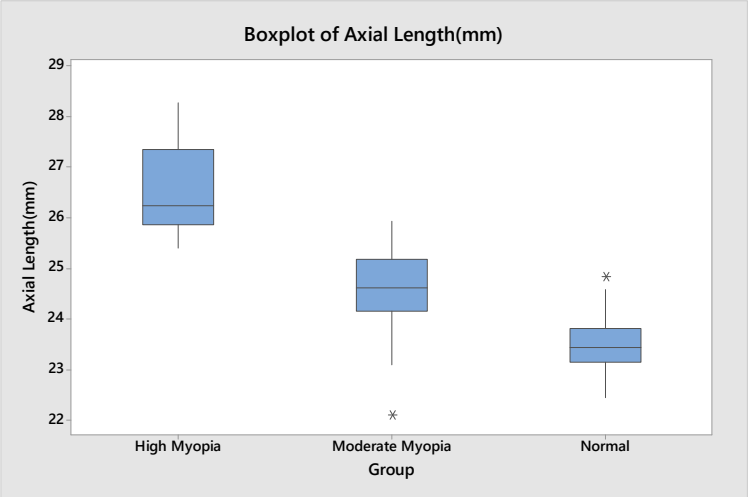
(b) Check the assumptions for a one-way ANOVA.

(a) Response is

Factor is

Levels are

Before an ANOVA, it is always a good idea to examine the data using graphical and numerical summaries.



Descriptive Statistics: Axial Length(mm)

Statistics

Variable	Group	N	Mean	StDev
Axial Length(mm)	High Myopia	14	26.505	0.861
	Moderate Myopia	28	24.583	0.859
	Normal	22	23.445	0.559

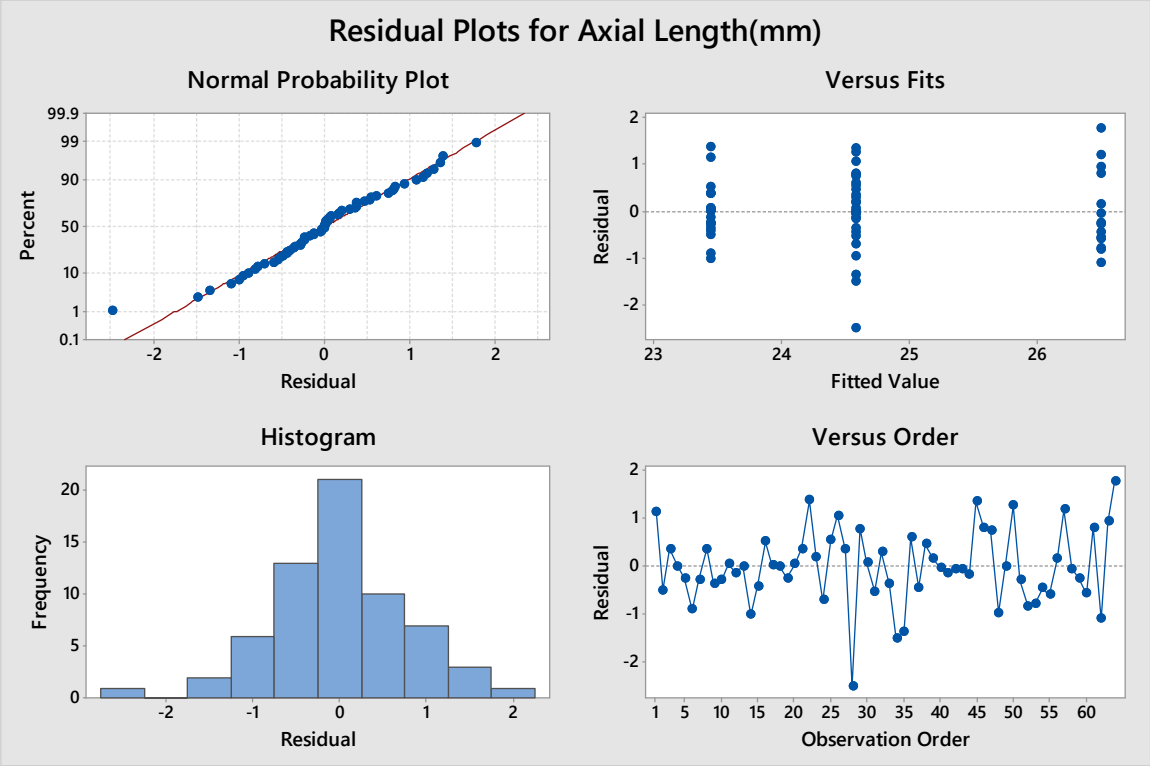
It appears here that subjects in the *High Myopia* group have the highest mean axial length with $\overline{x}_{High\ Myopia} = 26.505$ mm and those in the *Normal* group have the lowest mean with $\overline{x}_{Normal} = 23.445$ mm. Those in the *Moderate* group appear to have a mean in between. What can we conclude, albeit preliminary, about the differences in mean axial length of the eyes among these three groups in the population?

(b) Checking the assumptions for an ANOVA as follows:

- We view the three groups of subjects as independent random samples from three different populations.
- All three boxplots above appear fairly symmetrical so we assume the normality assumption.
  (Alternatively, produce residual plots in Minitab. See figure below.)
- $\dfrac{\text{largest sample SD}}{\text{smallest sample SD}} = \dfrac{0.861}{0.559} = 1.54 < 2$

  So, we assume that the variability in *Axial Length* across all three populations are approximately equal.

The assumptions of ANOVA are satisfied.



3. Comparing Means Using Variances

ANOVA makes use of a comparison of two different sources of variability:

- The average variation of observations within each group from the group mean. This is called the **within group variability**.
- The average variation of the group means from the overall mean (or grand mean). This is called the **between group variability**.

ANOVA then uses the F-test statistic: $F = \dfrac{\text{average varibility } \textit{between} \text{ groups}}{\text{average varibility } \textit{within} \text{ groups}}$

Let us, for convenience, consider only two populations to illustrate the concept. Let's simply call them population 1 and population 2, with means μ_1 and μ_2 respectively.
Suppose independent random samples, sample 1 and sample 2, are taken from each of these populations, giving means $\bar{x}_1 = 25$ and $\bar{x}_2 = 30$ (see *Figure 1*). Can we reasonably conclude that $\mu_1 \neq \mu_2$?

Next, suppose another two random samples, sample 3 and sample 4, are taken from population 3 and population 4, which have means μ_3 and μ_4 respectively. These two samples give $\bar{x}_3 = 25$ and $\bar{x}_4 = 30$ (see *Figure 2*). Again, can we infer that $\mu_3 \neq \mu_4$?

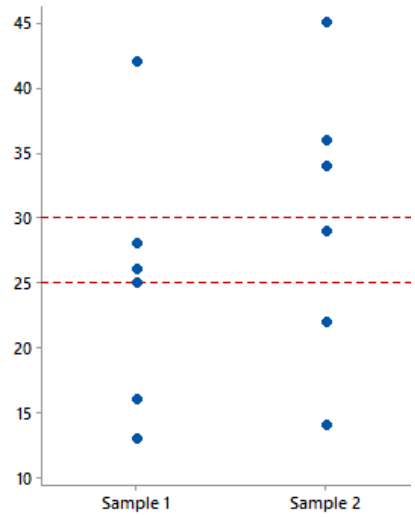


Figure 1: Comparing variability among groups with variability within groups

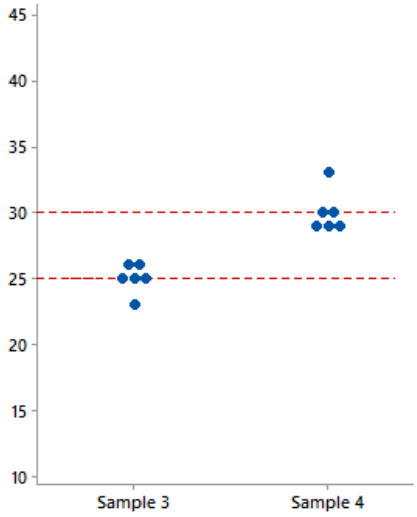


Figure 2: Comparing variability among groups with variability within groups

The sample means are the same in both figures, so the variability *between* groups is the same. But *Figure 1* shows relatively larger variability *within* the groups, while *Figure 2* shows smaller variability *within* the groups.
This suggests that the differences among the sample means in *Figure 1* could have occurred randomly due to chance. Thus, it is likely that $\mu_1 = \mu_2$.

However, *Figure 2* seems to suggest that the population means might differ. That is, $\mu_3 \neq \mu_4$.

This is the idea behind ANOVA: to assess whether several populations all have the same mean, we compare the variation *between* the means of several groups with the variation *within* groups.

4. The ANOVA Model

The Model

In a one-way ANOVA model, we take random samples of size n_i from each of the k different populations. Typical data for one-way ANOVA are as follows:

Group	Size	Sample Data	Group Sample Mean
1	n_1	$y_{11}, y_{12}, \cdots, y_{1n_1}$	$\overline{y}_1$
2	n_2	$y_{21}, y_{22}, \cdots, y_{2n_2}$	$\overline{y}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	n_k	$y_{k1}, y_{k2}, \cdots, y_{kn_k}$	$\overline{y}_k$
	Total = N		Overall mean = $\overline{y}_{\cdot\cdot}$

Note that when the sample sizes n_i are all the same, the experiment is said to be **balanced**, otherwise it is **unbalanced**.

The observations y_{ij} in the table above can be described by the one-way ANOVA model given by:

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{for } i = 1, 2, \ldots, k \quad \text{and} \quad j = 1, 2, \ldots, n_i$$

Here, y_{ij} represents the jth observation from group i. That is, the first digit of subscript indicates the group and the second digit indicates the subject in that group.
Then, ε_{ij} represents the random variation of the observations.

Example 4: Refer to *Case Study*.
What do the symbols/notations represent?

$k =$

$\mu_1 =$ $n_1 =$

$\mu_2 =$ $n_2 =$

$\mu_3 =$ $n_3 =$

$y_{11} =$, $y_{12} =$, $y_{13} =$, $\cdots$

$y_{21} =$, $y_{22} =$, $y_{23} =$, $\cdots$

$y_{31} =$, $y_{32} =$, $y_{33} =$, $\cdots$

The Hypotheses

The null and alternative hypotheses for a one-way ANOVA are:

$$\mathrm{H}_0 : \ \mu_1 = \mu_2 = \cdots = \mu_k$$
$$\mathrm{H}_1 : \ \text{At least one } \mu \text{ is different from the rest}$$

If the null hypotheses is true, that is, $\mu_1 = \mu_2 = \cdots = \mu_k = \mu$, then each observation y_{ij} consists of the overall mean μ plus a random error component ε_{ij}.
This is equivalent to saying that all *N* observations are taken from a single normal distribution with mean μ and variance σ^2.
Therefore, if the null hypothesis is true, changing the levels of the factor has no effect on the mean response.

The ANOVA Table

The analysis of variance partitions the total variability in the sample data into two components as follows:

$$SST = SS_{Between} + SS_{Within}$$

where, $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{\cdot}\right)^2 =$ total sum of squares

$$SS_{Between} = \sum_{i=1}^{k} n_i\left(\overline{y}_i - \overline{y}_{\cdot}\right)^2 = \text{ between group sum of squares}$$

$$SS_{Within} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2 = \text{ error sum of squares}$$

The proportion $R^2 = \dfrac{SS_{Between}}{SS_{Total}}$ is called the **coefficient of determination**.
It tells us the proportion of the variation in response that is explained by the factor, whilst the remaining proportion of the variation is probably explained by other factors.
R^2 is usually expressed in percentage, and ranges in value between 0% and 100%.

The ANOVA table can then be constructed as follows:

Source of variation	Sum of squares (*SS*)	Degrees of freedom	Mean square (*MS*)
Between groups	$SS_{Between} = \sum_{i=1}^{k} n_i\left(\overline{y}_i - \overline{y}_{\cdot}\right)^2$	$df_{Between} = k-1$	$MS_{Between} = \dfrac{SS_{Between}}{df_{Between}}$
Within groups	$SS_{Within} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2$	$df_{Within} = N-k$	$MS_{Within} = \dfrac{SS_{Within}}{df_{Within}}$
Total	$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{\cdot}\right)^2$	$N-1$	

To compare the population means $\mu_1, \mu_2, \ldots, \mu_k$, the F-test statistic is used to compare the variation between groups with the variation within groups. That is,

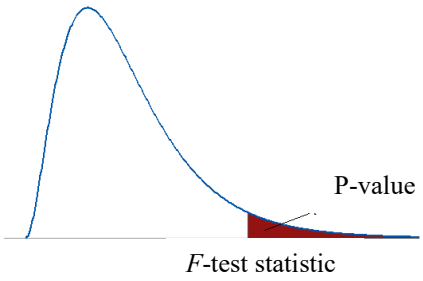
$$F = \frac{\text{mean square between groups}}{\text{mean square within groups}} = \frac{MS_{Between}}{MS_{Within}}$$

If the null hypothesis is true, we would expect the F-test statistic to be close to 1.

However, if the null hypothesis is not true, at least one of the sample means from a group will be far away from $\bar{y}_{\cdot}$, the overall mean of the entire data set.

This will cause $MS_{Between}$ to be large relative to MS_{Within}, which leads to an F-test statistic larger than 1. We reject the null hypothesis in favour of the alternative hypothesis if the F-test statistic is sufficiently large.

The P-value of the F test is the probability that a random variable having the $F(df_{Between}, df_{Within})$ distribution is greater than or equal to the calculated value of the F-test statistic, shown in the figure here:



Note that ANOVA is always an upper tail F test.

The computations needed for an ANOVA are generally tedious and lengthy, so we use statistical software to perform the calculations.

Example 5: Refer to *Case Study*.

(a) Set up the hypotheses.

(b) Read the output:

i. Verify that $SST = SS_{Between} + SS_{Within}$.

ii. Reconstruct the ANOVA table while verifying the numbers.

Source of variation	SS	df	MS
Between groups	$SS_{Between}$ =		
Within groups	SS_{Within} =		
Total	SS_T =		

iii. What is the F-test statistic? Can you compute it?

iv. Can you find the critical value? Hence, the critical region?

v. What is the P-value? Can you compute it?

vi. What is the conclusion?

vii. What is the coefficient of determination? Interpret it.

One-way ANOVA: Axial Length(mm) versus Group

One-way ANOVA: Axial Length(mm) versus Group

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Group	3	High Myopia, Moderate Myopia, Normal

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Group	2	80.18	40.0885	67.71	0.000
Error	61	36.11	0.5920		
Total	63	116.29			

The result of the significance test is given in this portion.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.769444	68.94%	67.93%	65.78%

This portion of the output gives descriptive statistics of the *Axial Length* for the three different groups. Besides the sample size, mean and standard deviation for each group, the 95% confidence interval for the population mean of each group is also given. Notice that all intervals do not overlap.

Means

Group	N	Mean	StDev	95% CI
High Myopia	14	26.505	0.861	(26.094, 26.916)
Moderate Myopia	28	24.583	0.859	(24.292, 24.874)
Normal	22	23.445	0.559	(23.117, 23.773)

Pooled StDev = 0.769444

A pooled estimate of the common standard deviation σ is also shown. The pooled standard deviation is a weighted average of the standard deviation of each group. It estimates a single standard deviation to represent the common standard deviation of all independent samples or groups in the study.

5. Multiple Comparison Tests

If the decision after an ANOVA is to not reject the null hypothesis, we conclude that the population means are indistinguishable, based on the data given. No further analysis is then necessary.

However, if the decision is to reject the null hypothesis, then we would like to know which pairs of means differ.

Example 6: Refer to *Case Study*.
State the number of possible pairs and list the pairs.

Multiple comparisons methods can be used to compare pairs of means:

$$H_0 : \mu_i = \mu_j \text{ or } \mu_i - \mu_j = 0$$
$$H_1 : \mu_i \neq \mu_j \text{ or } \mu_i - \mu_j \neq 0$$

for all means where $i \neq j$.

It is important to keep in mind that multiple comparisons methods are used only *after rejecting* the null hypothesis in ANOVA.

In performing a multiple comparisons procedure, we usually compute *t*-test statistics for all pairs of means using the formula:

$$t_{ij} = \frac{\overline{x}_i - \overline{x}_j}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

If t_{ij} falls into critical region corresponding to critical value t^* (d.f. = $n_i + n_j - 2$), then we conclude that the pair of population means is statistically different.

Otherwise, we conclude that the pair of population means is not statistically different, based on the data given.

Commonly used procedures for multiple comparisons include:

- Fisher's Least Significant Difference *t*-test
- Tukey's Honestly Significant Difference Multiple Comparisons Test
- Bonferroni Test

The above calculations are tedious so we use statistical software (e.g. Minitab) for the computations.

Example 7: Refer to *Case Study*.

Use Tukey's method to perform multiple comparisons test and read the output:

(a) What are the *t*-test statistics?

Moderate myopia vs. high myopia: test statistic = _____
normal vs. high myopia: test statistic = _____
normal vs. moderate myopia: test statistic = _____

(b) What are the P-values?

Mod. vs. high: df = _____, P-value ≈ _____
Normal vs. high: df = _____, P-value ≈ _____
Normal vs. mod.: df = _____, P-value ≈ _____

(c) What are the point estimates? What are the confidence intervals? What parameters do they estimate? What can you infer from the intervals?

Mod. vs. high: Point estimate for _____ is _____
95% CI for _____ is _____, which _____.
There is _____ difference in mean axial length of the eye between moderate myopia and high myopia groups.
Normal vs. high: Point estimate for _____ is _____
95% CI for _____ is _____, which _____.
There is _____ difference in mean axial length of the eye between normal and high myopia groups.
Normal vs. mod.: Point estimate for _____ is _____
95% CI for _____ is _____, which _____.
There is _____ difference in mean axial length of the eye between normal and moderate myopia groups.

(d) What is the conclusion?

Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Group	N	Mean	Grouping
High Myopia	14	26.505	A
Moderate Myopia	28	24.583	B
Normal	22	23.445	C

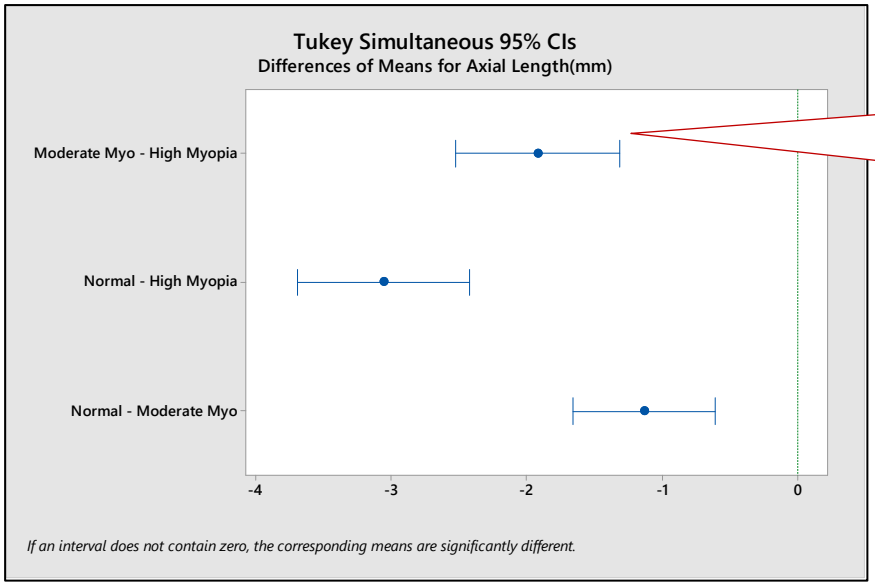
Means that do not share a letter are significantly different.

This portion in the output shows that each group is assigned a "Grouping" letter. Groups do not share a common letter significantly different means.

Tukey Simultaneous Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
Moderate Myo - High Myopia	-1.922	0.252	(-2.528, -1.317)	-7.63	0.000
Normal - High Myopia	-3.060	0.263	(-3.693, -2.428)	-11.63	0.000
Normal - Moderate Myo	-1.138	0.219	(-1.665, -0.611)	-5.19	0.000

Individual confidence level = 98.07%



The "dot" represents the difference of means and the "whiskers" represent the limits of the confidence interval. Note that it does not capture "0".

6. ANOVA vs. Multiple *t*-tests

Question: When we compare the means between three groups or more, why don't we just perform multiple independent two-sample *t*-tests?

One problem with this approach is that, as the number of independent groups increases, the number of two-sample *t*-tests also increases, leading to a higher probability of making a Type I error for the analysis.

For example: suppose we have 4 independent groups – A, B, C and D.
Then, the possible pairs to be compared are: _____
That will be ${}_4C_2 = 6$ independent *t*-tests that have to be conducted.
Suppose the probability of a making Type I error (α) in one such test is set at 0.05.
If we assume that the 6 independent *t*-tests are independent, then the *overall* probability of making at least a Type I error for the analysis is:

$$1-(0.95)^6 = 0.2649$$

Hence, multiple *t*-tests would lead to a higher overall Type I error rate, which is unacceptable. An ANOVA controls for this overall error rate, so that the Type I error remains at 5% and you can be more confident that any statistically significant result you find is not just due to running many tests.

TUTORIAL 7

1. A pharmaceutical company is studying the effectiveness of a new cream to relieve arthritis pain. The three preparations differ in the concentration of the active ingredients: T15 contains 15% active ingredients, T30 contains 30% active ingredients, and T45 contains 45% active ingredients. A sample of 80 patients are selected and randomly assigned to receive the three different preparations and a standard treatment, which serves as the control. The time (in minutes) until pain relief is recorded on each subject. An ANOVA was run in Minitab and the results are shown below.

One-way ANOVA: Control, T15, T30, T45

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	$\alpha = 0.05$

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Factor	4	Control, T15, T30, T45

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	3	372.7	124.246	22.86	0.000
Error	76	413.1	5.436		
Total	79	785.9			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.33153	47.43%	45.35%	41.75%

Means

Factor	N	Mean	StDev	95% CI
Control	20	16.350	2.796	(15.312, 17.388)
T15	20	18.200	2.546	(17.162, 19.238)
T30	20	15.250	2.173	(14.212, 16.288)
T45	20	12.250	1.650	(11.212, 13.288)

Pooled StDev = 2.33153

(a) State the response, factor, number of levels and levels.

(b) Is this a balanced or unbalanced design?

(c) Set up the null and alternative hypotheses.

(d) Identify the value of the test statistic with its degrees of freedom.

(e) Find the critical value at 0.05 significance level.

(f) Identify the *P*-value.

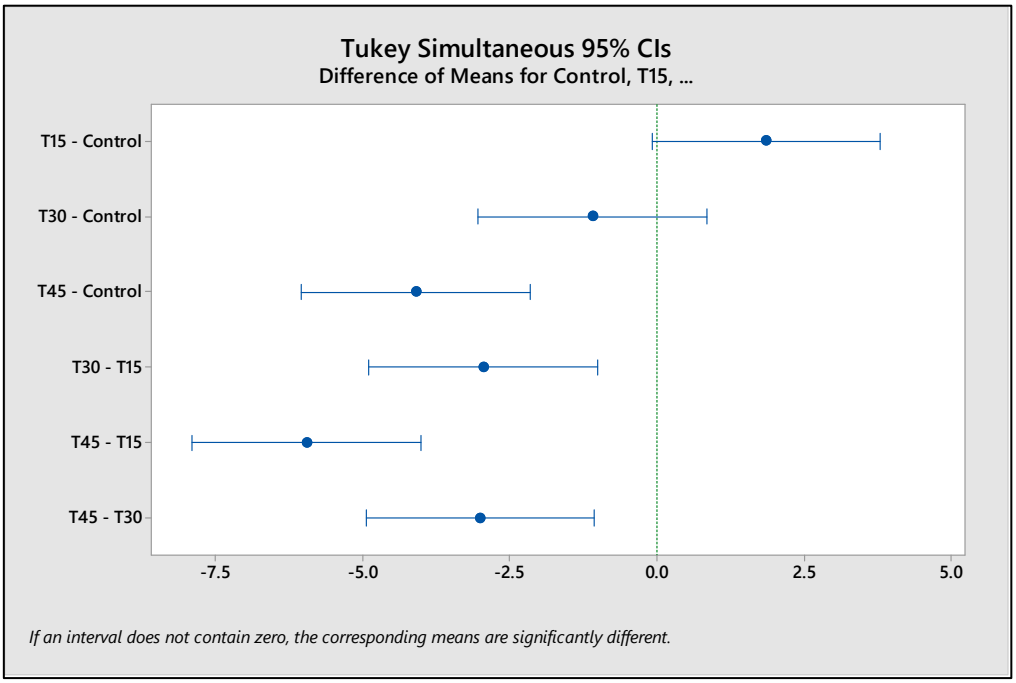
(g) What is the value of coefficient of determination? Interpret it.

(h) Based on the preceding results, what do you conclude about the effectiveness of the three preparations to relieve arthritis pain?

(i) The pharmaceutical company proceeded to conduct Tukey's multiple comparisons test in Minitab. A graph output is given below.

i. Explain why multiple comparisons test need to be conducted.

ii. Hence, which pairwise means differ?



2. Complete each ANOVA table below. Also determine the number of means being compared and summarize your conclusion.

(a)

Source	DF	SS	MS	F-Value	P-Value
Treatment	2	7.3884			
Error	15	4.1060			
Total					

(b)

Source	DF	SS	MS	F-Value	P-Value
Treatment	3	1092			
Error	38				
Total		4732			

ANSWERS

1. (a) Time until pain relief; cream; $k = 4$; control, T15, T30, T45
   (b) Balanced
   (c) $H_0 : \mu_{control} = \mu_{T15} = \mu_{T30} = \mu_{T45}$ vs. H_1 : At least one μ is different from the rest
   (d) $F = 22.86$; d.f. = 3, 76
   (e) 2.725
   (f) P-value < 0.001
   (g) $R^2 = 47.43\%$ of the variation in time until pain relief is explained by the cream type, whilst the remaining 52.57% is explained by other factors.
   (h) Reject H_0. There is very strong evidence that the mean time to pain relief differs in at least one treatment. T45 seems to offer the shortest time to pain relief among the four treatments.
   (i) i. Because H_0 was rejected, we need to know which pairs of means are significantly different.
   ii. The mean time to pain relief for T45 is shorter than those of the rest. Mean time to pain relief between T15 and T30 seems to differ too.

2. (a) $k = 3$ means are being compared.
   Since P-value = 0.00044 < 0.05, reject H_0 and conclude that at least one μ is different from the rest.
   (a) $k = 4$ means are being compared.
   Since P-value = 0.01776 < 0.05, reject H_0 and conclude that at least one μ is different from the rest.

Practical 7 : ANOVA

[Data can be downloaded from eSP *Practical* folder.]

Task 1

The financial structure of a firm refers to the way the firm's assets are divided by equity and debt, and the financial leverage refers to the percentage of assets financed by debt.
A researcher claims that financial leverage can be used to increase the rate of return on equity. The data below give the rates of return on equity using 3 different levels of financial leverage and a control level (zero debt) for 24 randomly selected firms.

Financial Leverage			
Control	Low	Medium	High
2.1	6.2	9.6	10.3
5.6	4.0	8.0	6.9
3.0	8.4	5.5	7.8
7.8	2.8	12.6	5.8
5.2	4.2	7.0	7.2
2.6	5.0	7.8	12.0

Use Minitab to perform an analysis of variance at the 0.05 level of significance.
Then, use Dunnett's test at the 0.05 level of significance to determine whether the mean rates of return on equity at the low, medium, and high levels of financial leverage are higher than at the control level.

Task 2

In an article "Shelf-Space Strategy in Retailing", published in *Proceedings: Southern Marketing Association*, the effect of shelf height on the supermarket sales of canned dog food is investigated. An experiment was conducted at a small supermarket for a period of 8 days on the sales of a single brand of dog food, involving three levels of shelf height: knee level, waist level and eye level. During each day, the shelf height of the canned dog food was randomly changed on three different occasions. The remaining sections of the gondola that housed the given brand were filled with a mixture of dog food brands that were both familiar and unfamiliar to customers in this particular geographic area. Sales, in hundreds of dollars, of this dog food per day for the three shelf heights are given, as follows:

Shelf Height		
Knee Level	Waist Level	Eye Level
77 82	88 94	85 85
86 78	93 90	87 81
81 86	91 94	80 79
77 81	90 87	87 93

Based on the data, is there a significant difference in the average daily sales of this dog food based on shelf height? Use a 0.05 level of significance.

(a) State the response, factor, number of levels and levels.

(b) Is this a balanced or unbalanced design?

(c) Set up the null and alternative hypotheses.

(d) Verify that the assumptions of an ANOVA are met.
(e) Identify the value of the test statistic with its degrees of freedom.
(f) Find the critical value at 0.05 significance level.
(g) Identify the *P*-value.
(h) What is the value of coefficient of determination? Interpret it.
(i) Based on the preceding results, what can you conclude?
(j) Conduct multiple comparisons test if needed. Explain the results.

Brief Answers

1. (a) P-value = 0.007, reject H_0.
   (b) Using Dunnett's test, the mean rates of return when financial leverage is set at the Medium and High levels appear to be higher than that for Control (P-values = 0.015 and 0.017 respectively) at the 0.05 significance level. The difference in mean return rates between Control and Low does not appear to be statistically significant (P-value = 0.902).
2. $F = 14.52$, d.f. = 2,21, P-value ≈ 0
   Sales from "waist-level" highest among all.

CHAPTER 8
CHI-SQUARE TEST

Learning Objectives:

1. *Identify statistical hypotheses (homogeneity & independence) to be tested.*
2. *Formulate correct null and alternative hypotheses.*
3. *Perform Chi-square test using statistical software.*
4. *Interpret P-value to make statistical inference.*
5. *Analyse case study involving comparing counts using the statistical problem-solving process.*

Content

1. Categorical Data

Suppose we have the following questions:

- Is smoker type associated with the occurrence of Age-related Macular Degeneration (AMD)?
- Is the distribution of students' choices of post-graduation activity (University, Work, Others) the same across four schools in a polytechnic (Business, Engineering, Life Sciences and Design)?
- Is the size of a tube of toothpaste (small, regular, large) purchased associated with the size of the buyer's household (small household of 1–2 persons, medium household of 3–4 persons, big household of 5–6 persons and extended household of 7 or more persons)?

What kind of data will we collect to answer these questions?

Questions asked	*What variables do we examine to answer the question?*	*What responses will we receive?*
Is smoker type associated to the occurrence of Age-related Macular Degeneration (AMD)?	Smoker type	❑ Smoker ❑ Non-smoker
	Occurrence of AMD	❑ Yes ❑ No
Is the distribution of students' choices of post-graduation activity the same across four schools in a polytechnic?	Post-graduation activity	❑ Work ❑ University ❑ Others
	School	❑ Business ❑ Engineering ❑ Life Sciences ❑ Design
Is the size of a tube of toothpaste purchased associated with the size of the buyer's household?		

Observe the type of data we may collect to answer the questions. Notice that they are non-numerical values that are descriptive in nature. They are called **categorical** or **qualitative** data.
Categorical data may be summarised by tallying the total number of occurrences. This tally is simply called **count**, or **frequency**.
The counts in categorical data are usually organised and presented in a **two-way table**, also known as the **contingency table**.

For example, the following tables show data collected to answer the corresponding question:

- **Is smoker type associated with the occurrence of AMD?**

  A sample of 900 elderly people was selected at random from a *single population* and classified accordingly to whether they were smokers or non-smokers, and whether there was evidence of the presence of AMD. The data obtained were summarized in the table as follows:

	Occurrence of AMD		
	Yes	**No**	**TOTAL**
Smokers	15	111	126
Non-smokers	69	705	774
TOTAL	84	816	900

  If there is no association, then we would *expect* that the proportion of occurrence of AMD should be the same regardless of smoker type. So do the *observed* counts match what are expected?

- **Is the distribution of students' choices of post-graduation activities the same across four schools in a polytechnic?**

  A sample of 1456 polytechnic students was selected from *different populations* classified according to the schools they were from, and what post-graduation activities they would be embarking on. The data obtained were summarized in the table as follows:

	School				
	Business	**Engineering**	**Life Sciences**	**Design**	**TOTAL**
University	209	198	177	101	685
Work	104	171	158	33	466
Others	135	115	39	16	305
TOTAL	448	484	374	150	1456

  If the distribution of post-graduation activities is the same across the four different schools, then we would *expect* that the number of students in each post-graduation activities to be equally distributed across. So do the *observed* counts match what are expected?

- **Is the size of a tube of toothpaste purchased associated with the size of the buyer's household?**

  A sample of 512 customers who purchased toothpaste were surveyed. These customers were selected from a *single* population, and were asked what size of tube of toothpaste was purchased, as well as the size of their household. The data obtained were summarized in the table as follows:

		Size of household				
		Small	**Medium**	**Large**	**Extended**	**TOTAL**
Size of toothpaste tube purchased	**Large**	23	116	78	43	260
	Regular	54	25	16	11	106
	Small	31	68	39	8	146
	TOTAL	108	209	133	62	512

  If there is no association, then we would *expect* that the proportion of each size of toothpaste tube purchased should be the same regardless of size of household. So do the *observed* counts match what are expected?

In all the scenarios described above, the observed counts are unlikely to match what are expected. Subsequently, we would want to know if the differences in the counts observed and expected are due to chance, or are larger than we would expect by chance.
To formally test this, we will perform a statistical test based on a distribution called the **Chi-Square distribution**.

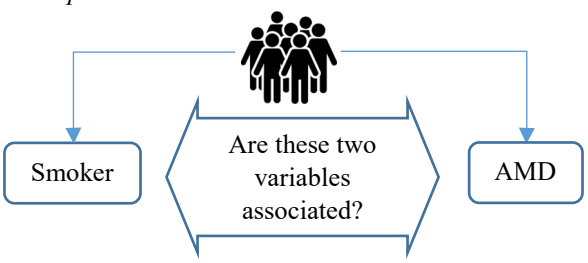
2. Hypotheses with Categorical Data

In previous chapters, quantitative data were taken by measuring a variable. Thus, hypothesis testing usually involves the mean(s) of that variable. Now that we are dealing with categorical data, what kind of hypotheses do we construct based on data taken from counting of variables?
There are two types of test for categorical data – for **independence** and for **homogeneity**.

Test for independence

Examines the distribution of counts for one group of individuals classified according to two categorical variables.

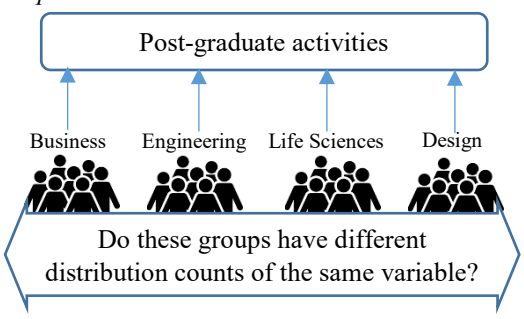
Example:



Test for homogeneity

Examines the distribution of counts for two or more groups of individuals in the same categorical variable.

Example:



The main difference between the two tests is the sampling method.

- In test for independence, one sample is taken, hence only the "grand" total number of samples is fixed. For instance, in the smoker/AMD scenario, a group of 900 elderly people was sampled.
  However, in test for homogeneity, multiple samples are taken, one from each population, hence the number of samples from each population is fixed. For instance, in the post-graduation activities scenario, 448 students were sampled from business school, 484 students were sampled from engineering school, 374 students were sampled from life sciences school and 150 students were sampled from designing school.
- As such, both variables in test for independence should have responses that cover all possible outcomes. For instance, in the smoker/AMD scenario, the possible outcomes for variable *smoker type* are "smoker" and "non-smoker"; the possible outcomes for variable *AMD* are "have" and "do not have". Both variables cover all possible outcomes.
  Whereas, only one of the variable in test for homogeneity should have responses that cover all possible outcomes. For instance, in the post-graduation activities scenario, the variable *activities* cover all possible outcomes, but the variable *school* does not. Hence sampling from each population has to be done for the variable *school*.

Consequently, the hypotheses for each test are written differently:

Test for independence	Test for homogeneity
H_0 : There is no association between the two variables (i.e. independent).	H_0 : Variable 1 is distributed the same across Variable 2 (i.e. homogeneous).
H_1 : There is an association between the two variables.	H_1 : Variable 1 is not distributed the same across Variable 2.

Despite the difference, we would be interested, in both tests, to test if the counts that we observed from the data are very different from what we will expect if null hypothesis is true. For example, we want to know if we can attribute the higher number of AMD patients to smokers; or could it be that the differences between the number of AMD patients in smoker and non-smoker groups occurred by chance?

Example 1: For each of the questions asked on pages 2 to 4, state the hypotheses and type of test.

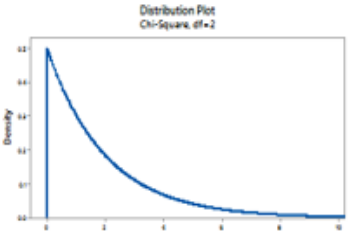
3. The Chi-Square Test Statistic

To test if the counts that we observed from the data are very different from what we will expect if there is no association between two variables (independence) or if the distribution across one variable is the same (homogeneity), we use a test based on the **Chi-square distribution**. [Note: *Chi* is pronounced as "kai"]

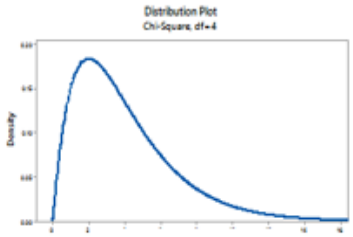
The Chi-Square Distribution

The Chi-square (χ^2) distribution is the distribution of the sum of squared standard Normal random variables. Hence, the χ^2 values begin at zero and are always non-negative. The graph of each χ^2 distribution is usually non-symmetrical and positively skewed; its shape will depend on its degree of freedom. As the degree of freedom increases, the χ^2 distribution becomes more and more symmetrical.

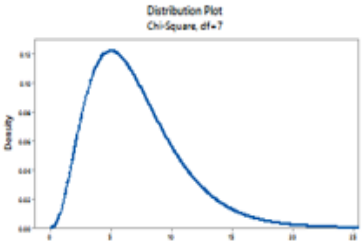
The following diagram shows the density curves of three Chi-Square distributions:



$\chi^2(df=2)$



$\chi^2(df=4)$



$\chi^2(df=7)$

The Chi-Square Test Statistic

Assuming a contingency table with *r* rows and *c* columns, then total number of cells = _____.

In each cell, there will be an **observed count** (or frequency) which has been tallied. We denote this with '*o*'.

Assuming independence or homogeneity, there will be an **expected count** (or frequency) in each cell. We denote this with '*e*'. How is *e* computed?

Let us illustrate with a 2-by-2 contingency table on gender vs. smoking:

	Male	Female	Total
Smoker	*e* = ?	*e* = ?	70
Non-smoker	*e* = ?	*e* = ?	130
Total	150	50	200

If smoking and gender are independent events, then:

P(Male Smoker) = P(Smoker)×P(Male) = ——×——

So, number of Male Smoker = 200×P(Male Smoker) =

Thus, in generalising, *e* in each cell can be computed as such:

$$e = \frac{(\text{row total}) \cdot (\text{column total})}{(\text{grand total})}$$

To test the hypothesis of independence or homogeneity, we compute the χ^2 test statistic:

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \cdots + \frac{(o_{rc} - e_{rc})^2}{e_{rc}}$$

Note:

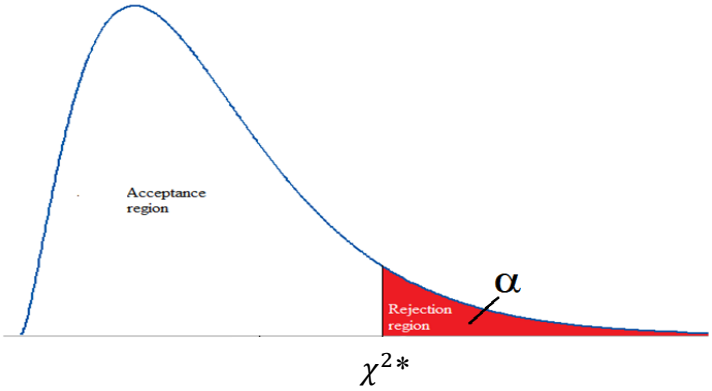
- The expected counts in any row or column add up to the appropriate *marginal* total. This allows us to compute one expected frequency in, say, the top row and then find the rest by subtraction.
- The expected count in each cell must be at least 5. If this condition is violated, then there are other tests that could be more appropriate (e.g. Fisher's exact test).

Critical Region & P-Value

The number of degrees of freedom for the χ^2 distribution is given by:

$$df = (r-1)(c-1)$$

The critical value χ^{2*} can be found from statistical software.



If χ^2 test statistic falls into the rejection region, then we will reject H_0.

Alternatively, if P-value = $P(\chi^2 > \chi^{2*})$ is smaller than significance level α, then we will reject H_0.

Notice that Chi-square test is always an upper tail test.

Case Study 1: Malaria

Reference: "Inference for Two-Way Tables: Against All Odds—Inside Statistics." Films Media Group, 2013,
https://eliser.lib.sp.edu.sg/ezp?rdURL=http://fod.infobase.com/PortalPlaylists.aspx?wID=151497&xtid=111548

One of the tropical diseases that kills and sickens millions of people every year, especially children is malaria. With malaria, we already know about an important source of resistance to the disease – it is a genetic mutation that is better known for the harm it does than the good. In fact, we have already encountered it in "*Against All Odds*" in the chapter on Binomial distribution. It is *sickle cell anaemia*.

Most kids in the United States who are sickle cell carriers are protected against malaria. It is this protective effect that is responsible for the sickle cell mutation becoming so prevalent. Data from research on 630 genes in these sick kids showed 7 instances of the HbS sickle cell haemoglobin gene, while the other 623 were the normal HbA haemoglobin gene.

The idea was to quantify whether children who came down with malaria were <u>less likely</u> to have inherited the protective sickle cell version of the gene rather than the normal version as compared to the general population.

The two-way table tallied the data collected. Observe that HbS was inherited by kids who caught malaria only 1.11% of the time, whereas in the control group, made of the general population, HbS was inherited 8.66% of the time.

Formulating question

Is the difference in proportion of kids who inherited HbS larger than we would expect by chance?
Is the difference statistically significant?

Basically, we want to know if there is sufficient evidence that the status of the two variables – malaria and HbS/HbA – are linked.

H_0 :

H_1 :

Type of test:

Collecting data

The data were obtained and tabulated as follows:

	HbA Susceptible	HbS Protective	Total
Malaria	623	7	630
Control	1065	101	1166
Total			

No. of rows, r = No. of columns, c =

Analysing data

Minitab output:

Chi-Square Test for Association: C1, Worksheet columns

Can you derive these numbers?

Rows: C1 Columns: Worksheet columns

	HbA	HbS	All
Malaria	623	7	630
	592.1	37.9	
control	1065	101	1166
	1095.9	70.1	
All	1688	108	1796

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	41.263	1	0.000
Likelihood Ratio	52.546	1	0.000

**Notice that all the expected counts are at least 5.*

If there is *no association* between having the HbS gene and catching malaria, we will expect to find 37.9 HbS genes in the children with malaria. But in reality, there are only 7 HbS genes in that group.

Is the difference between 7 and 37.9 large enough to tell us that there is an association between our two categorical variables? The next step in our analysis is to use the Chi-Square test to figure out if that difference is significant.

From the Minitab output:

Can you compute them?

χ^2 test statistic =

d.f. =

P-value =

(*Try using critical value method.*)

Interpreting results

So, we reject H_0, and conclude that there is very strong evidence of an association between HbS gene and catching malaria.

This gives support to our research hypothesis that the HbS sickle cell variant of the haemoglobin gene does protect against malaria.

Case Study 2: Choice of Post-Graduation Activities

As part of improvement in students' learning experiences, institutions of higher learning (IHLs) usually gather information about their graduates through graduate survey. One of the survey items is students' choice of activity upon graduation, such as furthering studies in universities, joining the workforce or other activities including taking a gap year, travelling, etc. Such information will help IHLs review the relevance and quality of courses that is offered.

Formulating question / Generating hypothesis

Is the distribution of students' choices of post-graduation activities the same across four schools in a polytechnic?

H_0 : post-graduation activities are distributed the same across four schools

H_1 : post-graduation activities are not distributed the same across four schools

Type of test:

Collecting data

The following data were collected from a polytechnic and tabulated:

	School				
	Business	**Engineering**	**Life Sciences**	**Design**	**TOTAL**
University	209	198	177	101	685
Work	104	171	158	33	466
Others	135	115	39	16	305
TOTAL	448	484	374	150	1456

No. of rows, r = No. of columns, c =

Analysing data

Minitab output:

Chi-Square Test for Association: C1, Worksheet columns

Rows: C1 Columns: Worksheet columns

	Business	Engineering	LifeSciences	Design	All
University	209	198	177	101	685
	210.8	227.7	176.0	70.6	
Work	104	171	158	33	466
	143.4	154.9	119.7	48.0	
Others	135	115	39	16	305
	93.8	101.4	78.3	31.4	
All	448	484	374	150	1456

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	93.657	6	0.000
Likelihood Ratio	96.771	6	0.000

Interpreting results	Are all the expected counts at least 5? From the Minitab output: χ^2 test statistic = d.f. = P-value = (*Try using critical value method.*)

TUTORIAL 8

1. For each of the following scenarios, state whether Chi-square test can be used.
   If can, decide if the test is to test for independence or to test for homogeneity.
   If Chi-square test cannot be used, explain why.

   (a) A brokerage firm wants to find out if the type of account a customer has (Silver, Gold or Platinum) affects the mode of trade that the customer uses (in person, by phone or on the internet). The firm collects a random sample of trade made for its customers over the past year and performs a test.

   (b) The Academic Services Centre in a polytechnic wants to investigate whether the distribution of internship companies (Multinationals, Small-Medium Enterprises and Start-ups) is chosen by a somewhat equal number of students for three cohorts. The Centre samples data from each cohort and performs a test.

   (c) A salesman who is on the road visiting clients thinks that, on average, he drives the same distance each day of the week. He keeps track of his car mileage for several weeks and performs a test on the data collected.

2. An analyst at a local bank wonders if the age distribution of customers coming for service at his branch in town is the same as the branch located near the mall. He selects 100 transactions at random from each branch and researches the age information for the respective customers. Here is the data tabulated:

	Less than 30	**30 to 55**	**56 or older**	**Total**
In-town branch	20	40	40	100
Mall branch	30	50	20	100
Total	50	90	60	200

   (a) What are the null and alternative hypotheses?

   (b) What is the type of Chi-square test?

   (c) What is the expected counts for each cell if the null hypothesis is true?

   (d) Find χ^2 test statistic. What is the degree of freedom?

   (e) Find the P-value.

   (f) Based on parts (a) to (e), what can the analyst conclude?

3. In an experiment to study the association of hypertension with smoking habits, data were taken from a sample of 180 individuals and tabulated as follows:

	Non-Smokers	**Moderate Smokers**	**Heavy Smokers**
Hypertension	21	36	30
No Hypertension	48	26	19

Test the hypothesis that the presence of hypertension is associated with smoking habits. Use a 0.05 level of significance.

4. Is the size of a tube of toothpaste purchased associated with the number of persons in the buyer's household? A sample of 512 customers who purchased toothpaste were surveyed. These customers were selected from a *single* population, and were asked what size of tube of toothpaste was purchased, as well as the size of their household. The data obtained were summarized in the table as follows:

		Size of household				
		Small	**Medium**	**Large**	**Extended**	**TOTAL**
Size of toothpaste tube purchased	**Large**	23	116	78	43	260
	Regular	54	25	16	11	106
	Small	31	68	39	8	146
	TOTAL	108	209	133	62	512

Use a 0.05 level of significance to test whether the number of persons in the household is associated with the size of tube of toothpaste purchased.

5. Daily Bread delivers 100 fresh bread to three different types of supermarket daily – VillageGrocer which specialises in imported goods, BudgetSaver which specialises in local produce, and BulkCo which specialises in bulk purchases. Daily Bread wants to check if there is a difference in the amount of bread sold or unsold in these three different types of supermarkets. The contingency table for the collected data, sampled from each supermarket, is presented here:

	Types of supermarket		
	VillageGrocer	**BudgetSaver**	**BulkCo**
Unsold bread	12	8	5
Sold bread	88	92	95

At the 0.05 significance level, is there evidence to believe that there is a difference in the number of bread sold in the supermarkets?

ANSWERS

1. (a) Test for independence
   (b) Test for homogeneity
   (c) Cannot be performed. Data are not categorical but quantitative (continuous).

2. (a) H_0 : age distribution of the customers at the two branches are the same.
   H_1 : age distribution of the customers at the two branches are not the same.
   (b) Test for homogeneity
   (c) The expected counts are:

	Less than 30	30 to 55	56 or older
In-town branch	25	45	30
Mall branch	25	45	30

   (d) 9.788, 2
   (e) 0.0075
   (f) The analyst can conclude that the age distribution of the customers at the two branches are not the same.

3. Test statistic = 14.464, d.f. = 2, P-Value = 0.001.
   The presence of hypertension is associated to smoking habits.

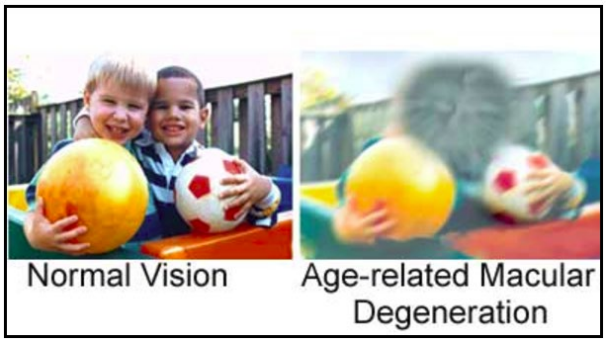
4. Test statistic = 89.34, d.f. = 6, P-Value = 0.000.
   The number of persons in the household is associated with the size of tube of toothpaste purchased.

5. Test statistic = 3.229, d.f. = 2, P-Value = 0.199.
   No evidence to believe that there is a difference in the number of bread sold in different types of supermarkets.

Practical 8 : Chi-Square Test

[Data can be downloaded from eSP *Practical* folder.]

Task 1: Smoking and Age-Related Macular Degeneration (AMD)



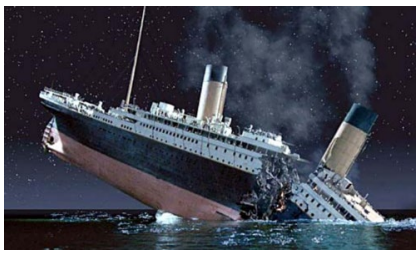
As the name implies, age-related macular degeneration (AMD) occurs when the macula degenerates with increasing age. The macula, located in the centre of the retina, provides information for fine, detailed vision when you look straight ahead. Damage to the macula results in problems with central vision.

When the macula fails to receive the nutrients it needs to survive, it degenerates. The symptoms of AMD includes difficulty in reading books, washed out colour and blind spots in the central visual field. (Source: *https://faculty.washington.edu/chudler/armd.html*)

The exact cause of AMD is not known. However, a group of scientists suspects that a person's smoking habit may be associated with AMD. Thus, they randomly sampled 900 subjects, of which, 84 are diagnosed as having AMD. Among those who have AMD, 15 are smokers, whereas among those who do not have AMD, 111 are smokers.

Conduct a Chi-square test for independence to check the scientists' claim.

Task 2: Titanic



The sinking of the Titanic is a famous event, and new books are still being published about it. Many well-known facts, from the proportions of first-class passengers to the "women and children first" policy, and the fact that the policy was not entirely successful in saving the women and children in the third class are discussed in great lengths. Based on the data provided, is there evidence to believe that passengers from different classes of cabins (First-, Second- and Third-class) have different survival status? Conduct a Chi-square test for homogeneity. Try to recode the data before you analyse.

Data code:

```
Class    (0 = Crew, 1 = First, 2 = Second, 3 = Third)
Age      (1 = Adult, 0 = Child)
Sex      (1 = Male, 0 = Female)
Survived (1 = Alive, 0 = Dead)
```

(Source: *https://ww2.amstat.org/publications/jse/datasets/titanic.txt*)

Brief Answers

(1) P-value = 0.285, do not reject H_0. (2) P-value < 0.001, reject H_0.

CHAPTER 9
SIMPLE LINEAR REGRESSION

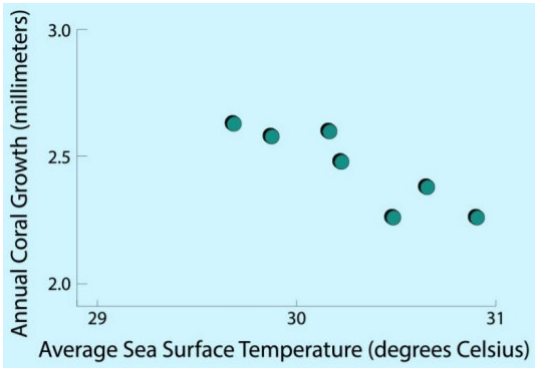
Learning Objectives:

1. *Identify explanatory and response variables in a regression problem statement.*
2. *Fit an estimated regression equation to a set of sample data based on least-squares method.*
3. *Describe the form of a simple linear regression equation.*
4. *Compute parameter estimates for a simple linear regression equation.*
5. *Use a regression equation to make predictions.*
6. *Interpret the marginal change in the dependent variable for a unit change in the independent variable.*
7. *Interpret the sample correlation coefficient, r, and the coefficient of determination, R^2.*
8. *Test hypothesis and make inferences about the slope and correlation coefficient of the simple regression model.*

Content

1. Fitting Lines to Data

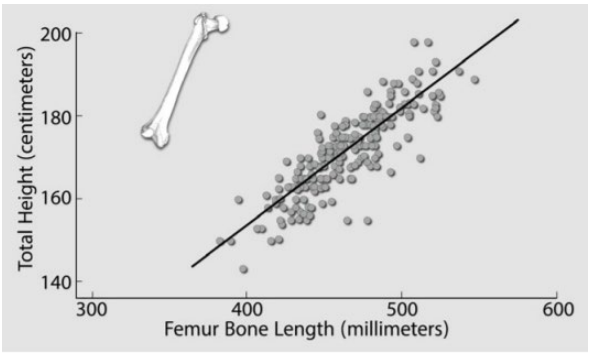
Scatterplot is a great way to visualize the **relationship** between two **quantitative** variables. Here are two such plots:



This scatterplot shows that as average sea surface temperature goes up, annual coral growth goes down.

Thus, a negative relationship is exhibited.

Note that temperature "explains" growth.



This scatterplot shows that as a person's femur bone length increases, so does total height.

Thus, a positive relationship is exhibited.

Note that bone length "explains" height.

Recall from *Descriptive Statistics: Analysing Relationships* (Chapter 1) that **explanatory variable** (or independent variable or predictor) always goes on the horizontal axis, and the **response variable** (or dependent variable or outcome variable) always goes on the vertical axis.

So, in the 1st scatterplot: Explanatory variable → average sea surface temperature
Response variable → annual coral growth

And, in the 2nd scatterplot: Explanatory variable →
Response variable →

In both the scatterplots, the data points appear to fall pretty much along a straight line, indicating a **linear** relationship between the two quantitative variables. This line that describes how the response variable changes with the explanatory variable is called a **regression line**.

Recall that any straight line can be described by the equation: $y = a + bx$

- "a" is the y-intercept. It is the y-value corresponding to $x = 0$.
- "b" is the slope (or gradient) of the line. It measures how much y changes when x increases by one unit.
- "x" and "y" are the data points that are plotted.

To figure out how best to fit a regression line to the data points, a statistical technique called **least-squares method** is employed. Then, we can use this best-fit line to make predictions.

Case Study 1: Forecasting Water Supply

Reference: "*Fitting Lines to Data: Against All Odds—Inside Statistics* (2013)". Films Media Group, https://eliser.lib.sp.edu.sg/ezp?rdURL=http://fod.infobase.com/PortalPlaylists.aspx?wID=151497&xtid=111530

Most of Colorado's water comes from melting mountain snow in the spring. The Colorado Climate Centre needs to forecast the state's seasonal water supply. Farmers, city planners and businesses, all need to know how much water is going to be available each year so they can plan accordingly. Climatologist Nolan Doesken introduces an important question for Colorado: How can we predict the water supply we are going to have as far ahead of time as possible?

To answer this question, climatologists have developed a model based on two types of data: the amount of winter snowpack in the high elevations and the resulting volume of water that flows out of the mountains throughout the summer. During the winter, Colorado's Natural Resources Conservation Service heads into the Rockies to collect data on the snowpack. Later, when the snowpack starts to melt, data related to the volume of water runoff are collected.

Are the variables/data of quantitative type?

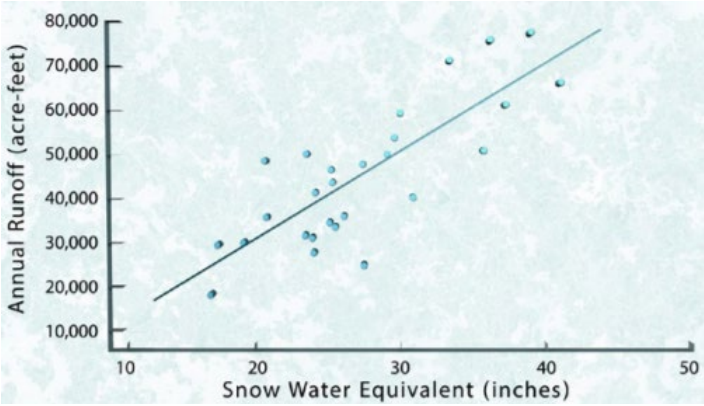
What is the explanatory variable?

What is the response variable?

The scatterplot on the right presents the data collected over many years.

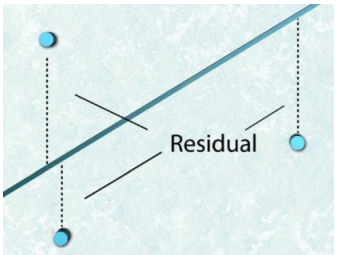
The scatterplot shows a strong positive linear relationship between amount of snowpack and volume of water runoff.

A line has been drawn to summarise this relationship.



In the real world, all the data points will not fall exactly on a line. So, we need a technique to determine the regression line that minimizes the vertical distances of the data points from the line.

To see how this can be done, let us zoom in on 3 data points:



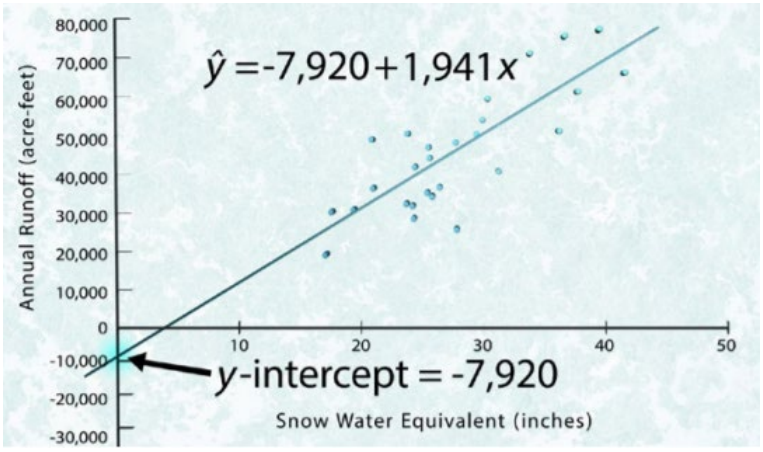
The vertical distance of a data point from the line is called **residual** (or **error**), denoted by e.

$$e = y - \hat{y}$$

where y is the y-value of the data point,
and $\hat{y}$ is the y-value of the line corresponding to the data point.

If the data point is above the line, then $e > 0$.
Else, if data point is below the line, then $e < 0$.
So the residuals are squared, and then sum together.
The best-fit line is the one where the sum of squared residuals (**SSE**) is the smallest. This is known as the **least-squares method**.

In the Colorado water supply data, this scatterplot shows both the graph of the least-squares regression line and its **regression equation**.



- $\hat{y}$ gives the *predicted* value of y, not the *measured* value from the data set.
- $b =$
  This means that for every 1 inch increase in snowpack, the runoff increases by 1941 acre-feet on average.
- $a =$
  This means that when there is no snowpack, the runoff would be −7,920 acre-feet. However, this does not make sense. Note that we should not extrapolate the regression line too far outside the range of the observed data.
- To use the least-squares line to make a prediction:
  If the Rockies saw 30 inches of snowpack, then we can predict that _____acre-feet of water is going to flow into the water supply in spring.

The regression line worked well to predict the Colorado water supply because the relationship between snowpack and runoff is linear. If the relationship has a curved pattern instead, a straight regression line will not be a good fit.

One way to assess how well a regression line fits the data is to make a **residual plot**, where the residuals are plotted against the explanatory variable. If the dots in the residual plot appear randomly scattered with no strong pattern, then the regression line has nicely captured the pattern in the data and a linear model is a good choice to describe it.

In this module, we will use statistical software to compute the regression equation and other regression estimates, as well as produce graphs.

2. Hypothesis Test for Slope of Regression Line

Suppose we have two quantitative variables, explanatory variable X and response variable Y, which are related in a linear manner, represented by the model:

$$y_i = \alpha + \beta x_i + \varepsilon_i \text{ , where } \varepsilon_i \text{ is the error term}$$

This model is called the **Simple Linear Regression** model since it has only one explanatory variable.
If slope β is zero, then y will be unaffected by how x changes. As such, there is no linear relationship between X and Y.

Given sample data, least-squares method will fit the best straight line with the equation:

$$\hat{y}_i = a + bx_i$$

This regression equation can be used to predict response y, given x.
Furthermore, the slope b and intercept a of the sample regression line are actually estimates for the "true" slope β and "true" intercept α of the population regression line. Hence, we can use the sample slope to make inferences about the population slope. We ask this: If sample data suggest a linear relationship between two variables, how can we determine whether this happened by chance or whether linear relationship really exist?

The hypotheses are set up as follows:

- $H_0 : \beta = 0$

  The slope of the population regression line is not significantly different from 0. That is, there is **no significant linear relationship** between the explanatory and response variables.

- $H_1 : \beta \neq 0$

  The slope of the population regression line is significantly different from 0. That is, there is a **significant linear relationship** between the explanatory and response variables.

(Note that correlation does not imply causation.)
The test for β is a two tail t-test with $n - 2$ degree of freedom. The test statistic is given as:

$$t = \frac{b - \beta}{SE(b)} \text{ ,}$$

where b is the slope of the sample regression line
β is the value of the slope of the population regression line under the null hypothesis
$SE(b)$ is the Standard Error of the slope of the regression line

Equivalently, and only in Simple Linear Regression, we can test for β using an upper tail F-test, similar to ANOVA. This is also known as the ***F*-test for lack of fit**.

We will use statistical software to perform the analyses.

Example 1: The local ice cream shop *Island Cream* keeps track of how much ice cream it sells versus the temperature on that day for the past 12 days. Using the data collected, *Island Cream* fitted a regression model with ice cream sales as the response and temperature of the day as the predictor:

Regression Analysis: Sales versus Temperature

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	160219	160219	110.22	0.000
Temperature	1	160219	160219	110.22	0.000
Error	10	14536	1454		
Total	11	174755			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
38.1265	91.68%	90.85%	88.15%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-159.5	54.6	-2.92	0.015	
Temperature	30.09	2.87	10.50	0.000	1.00

Regression Equation

Sales = -159.5 + 30.09 Temperature

Fits and Diagnostics for Unusual Observations

Obs	Sales	Fit	Resid	Std Resid	
11	445.0	520.5	-75.5	-2.17	R

R Large residual

Is there evidence of a linear relationship between temperature and sales?

3. Coefficient of Determination

How well does the regression model fit the data? Hence, how useful is the regression equation in making predictions?
To answer this, we need to measure the proportion of variation in response that is explained by the fitted model. This proportion is called the **coefficient of determination**, R^2.
The higher R^2 is, the better the fit the model is, and the more useful the equation will be in making predictions.

The total variation in responses is found by summing up all squared-deviations of responses from the mean value. This is known as **total sum of squares**, or simply, **SST**.

$$\text{SST} = \sum(y_i - \bar{y})^2$$

In linear regression, the total variation in the responses can come from two different sources:

- The variation in the responses that can be explained by the fitted model.
  This is found by summing up all squared-deviations of predicted values from the mean.
  This is known as **regression sum of squares**, or simply, **SSR**.

$$\text{SSR} = \sum(\hat{y}_i - \bar{y})^2$$

- The variation in the responses that cannot be explained by the fitted model.
  This is found by summing up all squared-residuals, that is, **SSE**.

$$\text{SSE} = \sum(y_i - \hat{y}_{i)})^2$$

So, we have that: $\text{SST} = \text{SSR} + \text{SSE}$

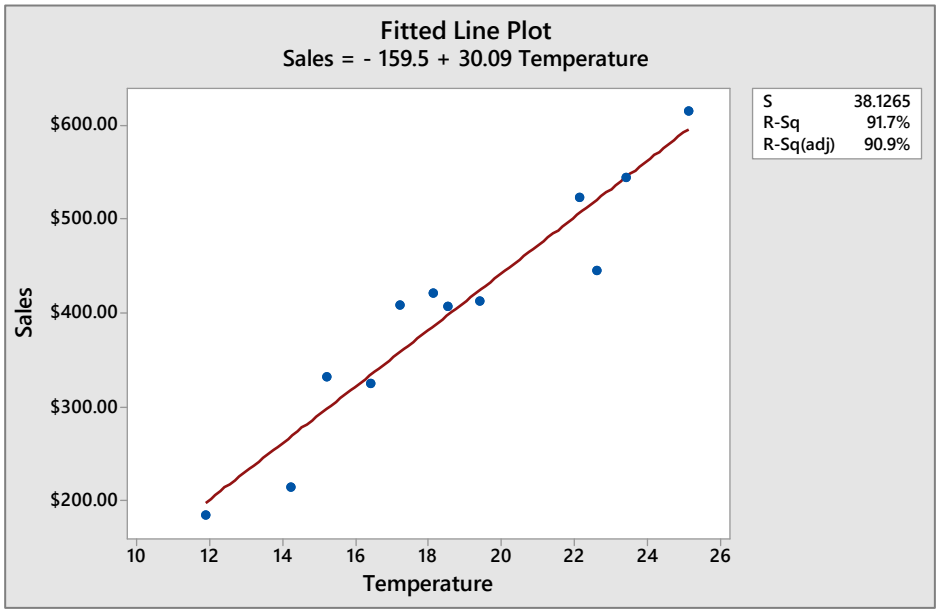
And, $$R^2 = \frac{\text{SSR}}{\text{SST}}$$

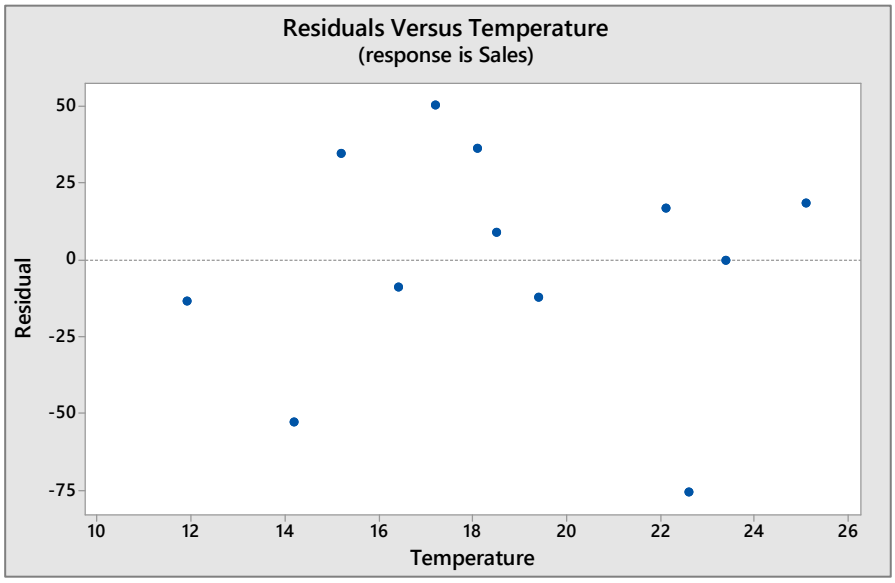
Note that R^2 is usually expressed as a percentage and it takes on a value between 0% and 100%.

Relationship between correlation coefficient, r, and coefficient of determination, R^2:

- Numerically, the coefficient of determination is equal to the square of the correlation coefficient.
- However, conceptually, r measures the strength of the linear relationship between two quantitative variables, while R^2 measures quality of fit of regression model.
- r is usually expressed as a number between -1 and 1, whereas R^2 is usually expressed as a percentage between 0% and 100%.

Example 2: Using the data collected, the local ice cream shop *Island Cream* puts the data through a statistical software and produced two graphs – a scatterplot with regression equation and a residual plot:





Read the graphs and answer the following questions:

(a) What is the regression equation? Interpret the coefficients.

(b) Is the regression model a good fit?

(c) If the model is useful for making predictions, predict the following sales. If prediction is not reliable, explain why.

  (i) What is the predicted sales on a day when temperature is 23°C?

  (ii) What is the predicted sales on a day when temperature is 34°C?

4. Assumptions

Certain assumptions have to be met in order to make inferences about the regression model. These assumptions, simplified, are:

1. **Linearity**: There are constants α and β such that the predicted (or mean) value of y in the underlying population is related to x by $E(y) = \alpha + \beta x$ for each value x of the predictor variable.
2. **Normality**: The distribution of y for a given value of x is normal. This is equivalent to saying that, for a given value of x, the residuals are normally distributed with mean zero.
3. **Constant standard deviation**: The variability about the regression line is the same for all values of x, with constant standard deviation.

5. Hypothesis Test for Population Correlation Coefficient

The correlation coefficient tells us about the strength of the linear relationship between two quantitative variables. A correlation coefficient of zero indicates no linear relationship between the two variables. Hence, we can conduct a hypothesis test on the correlation coefficient to check if it is "close to 0" or "significantly different from 0". We infer this from the sample correlation coefficient and the sample size.

The population parameter and sample statistic are:

ρ = population correlation coefficient

r = sample correlation coefficient

The population correlation coefficient is unknown but fixed. We will attempt to estimate it with the sample correlation coefficient, which is computed from sample data and varies from sample to sample.

Since the correlation coefficient between X and Y, or between Y and X is the same, this hypothesis test can be used if it is not obvious which variable should be treated as explanatory or response.

The hypotheses are set up as follows:

- $H_0 : \rho = 0$

  The population correlation coefficient is not significantly different from 0. That is, there is **no significant linear relationship** (correlation) between the two variables.

- $H_1 : \rho \neq 0$

  The population correlation coefficient is significantly different from 0. That is, there is a **significant linear relationship** (correlation) between the two variables.

(Note that correlation does not imply causation.)

The test for ρ is a two tail t-test with $n - 2$ degree of freedom. The test statistic is given as:

$$t = \frac{r - 0}{SE(r)}$$

where r is the sample correlation coefficient
$SE(r)$ is the Standard Error of the sample correlation coefficient

The value of the test statistic, t, is shown in the calculation output along with the P-value. The test statistic t has the same sign as the correlation coefficient, r. The P-value is the combined area in both tails.

Finding the correlation coefficient and testing hypotheses involving the correlation coefficient involve a fair bit of calculations, so we usually rely on the statistical software to do the computations for us.

Example 3: Using the data collected, *Island Cream* produced the following software output on correlation coefficient between ice cream sales and temperature of the day:

Correlation: Temperature, Sales

Correlations

Pearson correlation	0.958
P-value	0.000

Test if there is a linear relationship between sales and temperature.

Case Study 2: Measuring Stock Market Risk

The S&P 500 is widely regarded as the best single gauge of large-cap U.S. equities. The index includes 500 leading companies and captures approximately 80% coverage of available market capitalization. It powers countless index mutual funds and exchange-traded fund, and has become synonymous with "the market".

Investors can look at several different measures of stock market performance for a gauge of performance and an indication of the overall economy. The stock market is one of many different factors that economists consider when they look at economic health. The most common measures of performance are the market indices, with the Dow Jones Industrial Average and the S&P 500 being the most popular.

Formulating questions

A trader has a portfolio consisting of 7 popular stocks, namely Microsoft, Exxon Mobil, Caterpillar, Johnson & Johnson, MacDonald's, Scandisk and Qualcomm.
Using S&P 500 index as a gauge, which stock is the most volatile?

Collecting data

The mean monthly returns of the 7 stocks in the US stock exchange and S&P 500 index are collected over a 12-month period.
Explanatory variable: S&P 500 index
Response variable: Mean monthly return

Analyzing data

7 regression equations relating the S&P500 index to the mean monthly return of each stock is computed. The betas (slope of the estimated regression equation) for the individual stocks can be obtained from the regression output.

Company	*Beta*
Microsoft	0.458
Exxon Mobil	0.731
Caterpillar	1.490
Johnson & Johnson	0.009
MacDonald's	1.500
Scandisk	2.600
Qualcomm	1.410

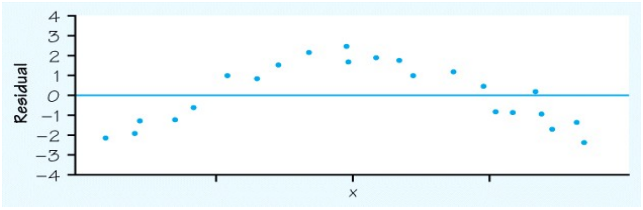
Interpreting results

The beta for the market as a whole is 1. So any stock with a beta greater than 1 will move up faster when the market goes up. Any stock with a beta less than 1 will not go down as fast as the market in period where the market declines.
We would expect Sandisk with a beta of 2.6 to benefit most from an up market, and lose most from a down market. Hence, Scandisk has the most risk.
Johnson & Johnson with a beta of 0.009 is least affected by the market.

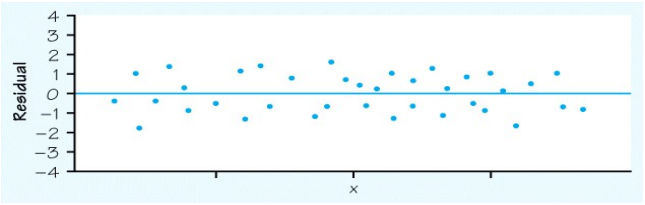
TUTORIAL 9

1. In compensation administration, factors affecting salary or benefits are often of interest to human resource (HR) professionals. To stay relevant, HR often has to decide on the best linear model to compensate staff fairly and consistently. Regression analyses allow HR professionals to consider all these factors in setting salaries and benefits levels.

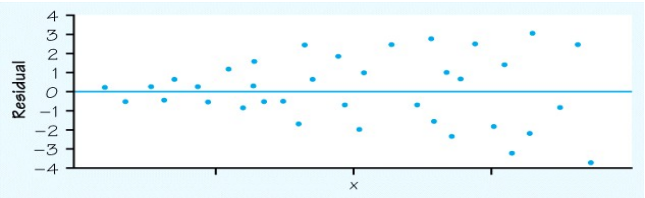
   Salaries of randomly selected data analysts are collected along with their relevant experience in years, performance levels and employees' ages. A linear regression model is fitted for each factor and the residual plots are produced as follows:



(a) Relevant experience in years



(b) Performance level



(c) Employer's age

   Which factor shows the best linear relationship with salary of data analyst in a pharmaceutical plant?

2. The National Directory of Magazines tracks the number of magazines published in Singapore each year. An analysis of data from 1988 to 2007 gives the following output.

Predictor	Coef	StDev	T	P
Constant	13549.9	2.731	7.79	0.000
Years	325.39	0.1950	10.0	0.000
S = 836.2	R-Sq = 84.8%	R-Sq (adj) = 80.6%		

   The dates were recorded as years since 1988. Thus, the year 1988 was recorded as year 0. A residual plot (not shown) showed no pattern. Read the output and answer the following questions:

   (a) What is the value of the slope of the least-squares regression line? Interpret the slope in this context.

(b) What is the value of the y-intercept of the least-squares regression line? Interpret the y-intercept in this context.

(c) Predict the number of magazines published in Singapore in 1999.

(d) What is the value of the correlation coefficient for number of magazines published in Singapore and years since 1988? Interpret the correlation coefficient in context.

(e) What is value of coefficient of determination? Interpret it in context.

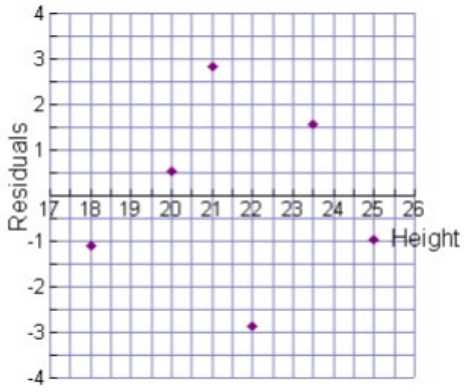
3. A researcher has a large number of data pairs (age, height) of human beings from birth to age 70 years. He computes a sample correlation coefficient, r.

(a) Would you expect r to be positive or negative? Why?

(b) What would you suggest to be a major problem with this approach?

(c) The researcher decides to use data only for adults aged between 21 to 60 years old to compute a sample correlation coefficient. What value of r should he expect?

4. The heights (in inches) and weights (in pounds) of six male Labrador Retrievers were measured. The height of a dog is measured at the shoulder. A simple linear regression analysis was done, and the residual plot and output are given below.



Predictor	Coef	StDev	T	P
Constant	−13.430	1.724	7.792	0.0000
Height	3.6956	0.4112	8.987	0.0004
S = 2.297	R-Sq = 95.3%	R-Sq (adj) = 90.6%		

(a) Is a linear line an appropriate model to fit these data? What information tells you so?

(b) Write the equation of the least-squares regression line. Identify the variables used in this equation. Interpret the coefficients.

(c) Is the regression model a good fit? Explain.

(d) Lucky, a male Labrador, was one of the dogs measured for this study. His height is 23.5 inches. Find Lucky's predicted weight and actual weight.

ANSWERS

1. The best linear model is *performance level*. The plot shows a uniform scatter of the points above and below the fitted line with no unusual individual observations.
   Relevant experience in years shows a curved pattern, the overall pattern of the data is not linear.
   Employer's age shows an increasing spread of salary about the line as age increases. Prediction for salary will be less accurate for larger values of ages.

2. (a) Slope is 325.39. For each year since 1988, the number of magazines published in Singapore increases by about 325 on average.
   (b) y-intercept is 13549.9. The predicted number of magazines published in Singapore in 1988 (year 0) is 13550 magazines.
   (c) 1999 is year 11. $\hat{y} = 13549.9 + 325.39(11) = 17129$.
   (d) Since the slope is positive, the correlation coefficient is the positive square root of 0.848, giving 0.921. This indicates a strong, positive, linear relationship between the number of magazines published in Singapore and years since 1988.
   (e) 84.8% of the variation in number of magazines published in Singapore can be explained by linear relationship between number of magazines and years since 1988.

3. (a) Positive, since in general people grow taller as they grow older.
   (b) The underlying data is not linear. During the first few years of life, height increases rapidly and irregularly. After teenage years, height is essentially constant. A correlation coefficient is a measure of the scatter about a straight line. A better plan would be to restrict the data set to children only.
   (c) $r \approx 0$. The heights of adult are generally constant; that is, it does not change with age.

4. (a) Yes, a linear model is appropriate. The residual plot shows no pattern and a test for slope shows that there is a relationship. H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$, where β is the slope of the weight vs. height graph. d.f. = 4, t = 8.987 ; P-value = 0.0004.
   (b) $\hat{y} = -13.43 + 3.6956x$, where x = height & $\hat{y}$ = predicted weight. For every 1 inch increase in height, the weight increases by 3.6956 pounds on average. When height is zero, weight is −13.43 pounds, which does not make sense as extrapolation is out of data range.
   (c) Yes, because R^2 = 95.3% is close to 100%. This means that 95.3% of the variation in weight can be explained by the fitted model.
   (d) $\hat{y} = -13.43 + 3.6956(23.5) = 73.4$ pounds. Residual (reading from graph) is approximately 1.6. Residual = $y - \hat{y} \rightarrow$ 1.6= $y - 73.4 \rightarrow y = 75$
   Actual weight is 75 pounds.

Practical 9 : Simple Linear Regression

[Data can be downloaded from eSP *Practical* folder.]

Task 1

The partial table below gives the BMI and total cholesterol of a sample of 10 participants who went for a health check in a clinic.

BMI	*Total Cholesterol*
27.6	221
26.3	180
⋮	⋮

(a) Construct a scatterplot for the data. Does the graph suggest anything about the nature of the relationship between BMI and total cholesterol?

(b) Find the least-squares regression line and R^2 between BMI and total cholesterol. Interpret them. Test the slope.

(c) Produce the residual plot of the fitted model. What do you observe?

Task 2

A pharmaceutical company wants to model the relationship between the salaries and years of working experience. The data below partially shows salary ($ per hour) and years of working experience for 35 staff.

Salary	Years
88	11
77.3	6
⋮	⋮

(a) Compute *r,* the correlation coefficient between salary and years. Test the hypothesis that the population correlation coefficient ρ is equal to zero. Use 0.05 level of significance.

(b) Fit a linear model on the data and state the regression equation. Is it a good fit? Test the hypothesis that the slope β is equal to zero. Use 0.05 level of significance.

Task 3

Is there statistically significant evidence of a linear relationship between price of flats (in US$1000) and distance of the flats to the nearest subway station (in km)? If so, is the relationship positive or negative? The data collected to answer the question are partially shown here:

Price of flat	Distance of flat to nearest subway station
211	2.5
167	3.9
⋮	⋮

Perform an appropriate hypothesis test at 0.05 level of significance.

Task 4

The following data give information on the ages (in years) and the number of breakdowns during the past month for a sample of seven machines at a large company.

Age	12	7	2	8	13	9	4
No. of breakdowns	10	5	1	4	12	7	2

(a) Find the least-squares line with age as the independent variable and the number of breakdowns as the dependent variable.

(b) If age is increased by 1 year, by how much would you predict the number of breakdowns to increase or decrease?

(c) Predict the number of breakdowns for a machine of age 5 years.

(d) Find r and R^2. Interpret them.

(e) Can the least-squares line be used to predict the number of breakdowns for a machine of age 15 years? If so, predict the number of breakdown. If not, explain why not.

(f) Is the least-squares line an appropriate model to use for these data?

Task 5

A researcher wants to investigate the relationship between calcium intake and knowledge about calcium in a population of tertiary students. A random sample of 20 students gave data as follows:

Respondent	Knowledge Score (out of 50)	Calcium Intake (mg/day)
1	10	450
2	42	1050
3	38	900
⋮	⋮	⋮

(Source: http://www.statstutor.ac.uk/resources/uploaded/coventrysimplelinearregression.pdf)

(a) Construct a scatterplot of calcium intake vs. knowledge score. Does the plot suggest anything about the nature of the relationship between these variables in the sample of students?

(b) Determine if any correlation exists between calcium intake and knowledge score. If so, describe the relationship.

In addition to the relationship between calcium intake and knowledge about calcium, the researcher also wants to know if knowledge about calcium can be used to predict calcium intake of the students in the population.

(c) Using calcium intake as the response and knowledge score as the predictor variable, compute the least-squares regression line. Interpret the estimated slope and intercept of the line.

(d) What is the estimated mean calcium intake for the population of students whose knowledge score is 20?

(e) Find coefficient of determination and interpret it.

Brief Answers

1. (a) Positive relationship.
   (b) $\widehat{\text{Total cholesterol}} = 1.1 + 7.65 \times BMI$. $R^2 = 77.86\%$. P-value = 0.001 < 5%
   (c) Residuals appear to be randomly scattered around zero with no strong pattern.
2. (a) $r = 0.912$, P-value < 0.001.
   (b) $\hat{y} = 60.7 + 2.169x$ $R^2 = 83.2\%$, P-value < 0.001
3. P-value < 0.001, negative
4. (a) $\hat{y} = -1.917 + 0.989x$
   (b) ≈ 1 breakdown
   (c) ≈ 3 breakdowns
   (d) $r = 0.97$; strong positive linear relationship. $R^2 = 94\%$; 94% of the variability in number of breakdowns can be explained by the fitted model.
   (e) Probably not, due to extrapolation.
   (f) Yes, a linear model is appropriate. H_0: $\beta = 0$, P-value < 0.001

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.917	0.975	-1.97	0.106	
Age	0.989	0.112	8.80	0.000	1.00

5. (a) Yes, as knowledge score increases, calcium intake increases too.
   (b) $r = 0.882$. P-value < 0.001. Strong, positive, linear relationship.
   (c) $\hat{y} = 373.7 + 13.90x$
   (d) 651.7
   (e) $R^2 = 77.8\%$