UNIVERSITI TUNKU ABDUL RAHMAN

ACADEMIC YEAR 2019/2020

JANUARY 2020 TRIMESTER

FINAL ASSESSMENT

ANSWER SCRIPT

Candidate is required to fill in ALL the information below:

Name : (as stated in Student Identity Card)	Ngu Yi Hui		
Faculty /Institute/ Centre:	FSc	Programme :	Statistical Computing And Operations Research
Index No. (in numbers) :	A00082DBSCF	Index No. (in words) :	A Zero Zero Zero Eight Two DBSCF
Course Code :	UDPS2033	Course Description :	Sample Survey And Sampling Techniques
Submission Date :	15th MAY 2020	Submission Time :	9am

QUESTION NUMBER	**FOR EXAMINER'S USE ONLY**	
	MARKS	
	Internal	**External**
Q1		
Q2		
Q3		
Q4		
TOTAL MARKS		

UTAR
UNIVERSITI TUNKU ABDUL RAHMAN

Wholly owned by UTAR Education Foundation
(Co. No. 578227-M)
DU012(A)

DECLARATION STATEMENT

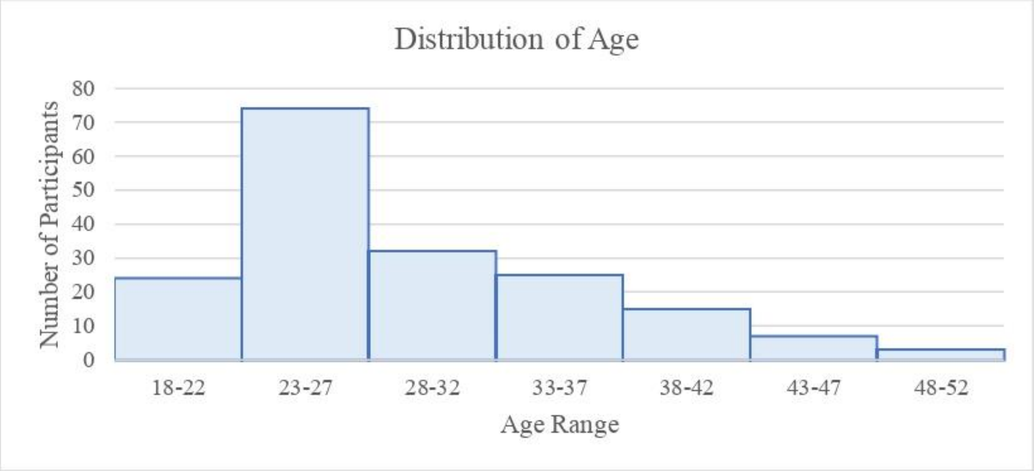
I, Ngu Yi Hui (Name), Student ID No. 18ADB01438, hereby solemnly and fully declare and confirm that during my programme of study at Universiti Tunku Abdul Rahman, I shall abide and comply with all the rules, regulations and lawful instructions of Universiti Tunku Abdul Rahman and endeavour at all times to uphold the good name of the University.

I hereby declare that my submission for this Final Assessment is based on my original work, not plagiarised from any source(s) except for citations and quotations which have been duly acknowledged. I am fully aware that students who are suspected of violating this pledge are liable to be referred to the Examination Disciplinary Committee of the University.

Programme:	Statistical Computing And Operations Research
(Digital) Signature:	HUI
Student's I.C / Passport No.:	991110-14-6378
Index No:	A00082DBSCF
Date of Submission:	15th MAY 2020

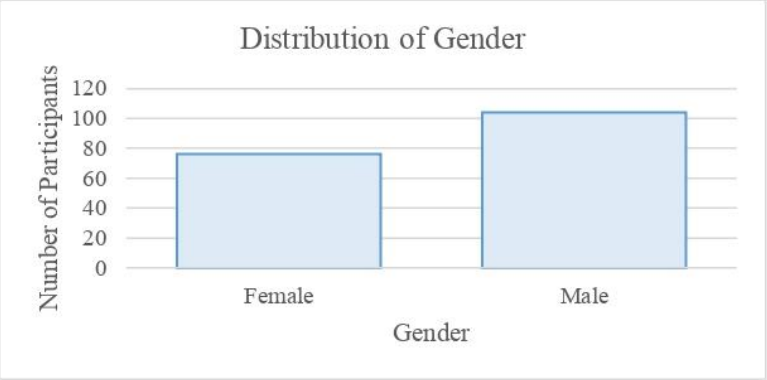
Question 1.

(i) Distribution of Age:



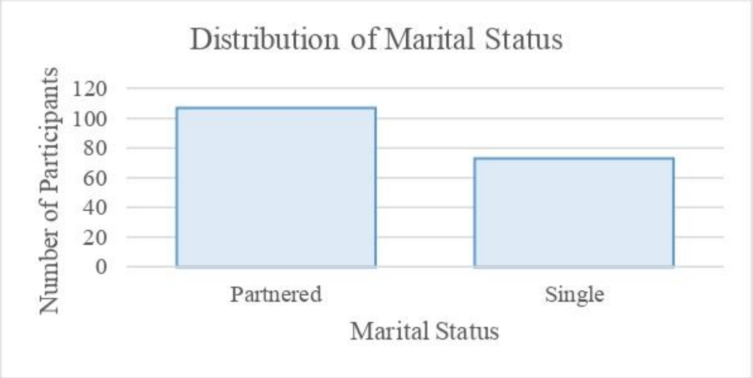
From the graph above, most buyers are aged 23 to 27 years old. The range for the age distribution is 18 to 50 years old. The mean of the age is 28.79, the median is 26, and the mode is 25. This shows that the distribution of age skewed to the right with a skewness of 0.982.

Distribution of Gender:



There are total 180 buyers for the company. 57.8% of them are male buyers, while 42.2% of them are female buyers. This shows that males are tend to buy the treadmills from the company.

Distribution of Marital Status:



From all 180 buyer, 59.4% of them are partnered, and 40.6% of them are single. This indicates the company has more customers from a married marital status.

Question 1.

(ii) Mean fitness level of buyers:

$$\bar{y} = \frac{\sum_{i=1}^{180} y_i}{n} = \frac{596}{180} = 3.3111$$

$$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{0.9194}{180}\left(\frac{2053-180}{2053}\right) = 0.00466$$

$$B = 2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{0.00466} = 0.1365$$

(iii) Mean income for buyers:

$$\bar{y} = \frac{\sum_{i=1}^{180} y_i}{n} = \frac{9669524}{180} = 53719.5778$$

$$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{272470624.1448}{180}\left(\frac{2053-180}{2053}\right) = 1381007.412$$

$$B = 2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{1381007.412} = 2350.3254$$

(iv) Difference in the mean income between female and male buyers:

Let $\bar{y}_1$ represents the mean income of females, $\bar{y}_2$ represents the mean income of males.

$$E(\bar{y}_1 - \bar{y}_2) = E(\bar{y}_1) - E(\bar{y}_2) = 49828.9079 - 56562.7596 = -6733.8517$$

$$\hat{V}(\bar{y}_1 - \bar{y}_2) = \hat{V}(\bar{y}_1) + \hat{V}(\bar{y}_2) = \frac{157695588.9}{76}\left(\frac{718.55-76}{718.55}\right) + \frac{339358580.6}{104}\left(\frac{1334.45-104}{1334.45}\right)$$

$$= 1855478.332 + 3008757.321 = 4864235.653$$

95% Confidence Interval $= (\bar{y}_1 - \bar{y}_2) \pm 2\sqrt{\hat{V}(\bar{y}_1) + \hat{V}(\bar{y}_2)} = -6733.8517 \pm 2\sqrt{4864235.653}$

$= (-11144.854, -2322.849)$

The 95% confidence interval of the difference does not include the value of zero.

Thus, there is a difference in the mean income between female and male buyers at 5% significance level.

(v) Proportions of females and males who rated their fitness level as satisfactory:

Let $\hat{p}_1$ represents the proportion of females who rated their fitness level as satisfactory,

$\hat{p}_2$ represents the proportion of males who rated their fitness level as satisfactory.

$$E(\hat{p}_1 - \hat{p}_2) = \hat{p}_1 - \hat{p}_2 = 0.1842 - 0.3942 = -0.21$$

$$\hat{V}(\hat{p}_1 - \hat{p}_2) = \hat{V}(\hat{p}_1) + \hat{V}(\hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2} = 0.00198 + 0.0023 = 0.0043$$

95% Confidence Interval $= (\hat{p}_1 - \hat{p}_2) \pm 2\sqrt{\hat{V}(\hat{p}_1) + \hat{V}(\hat{p}_2)} = -0.21 \pm 2\sqrt{0.0043}$

$= (-0.341, -0.079)$

The 95% confidence interval of the difference does not include the value of zero.

Thus, the proportions of females and males who rated their fitness level as satisfactory is different at 5% significance level.

Question 1.

(vi) Proportion of participants who rate their health as satisfactory over non-satisfactory among the males:

Let $\hat{p}_1$ represents the proportion of participants who rated their fitness level as satisfactory among males, $\hat{p}_2$ represents the proportion of participants who rated their fitness level as non-satisfactory among males.

$$E(\hat{p}_1 - \hat{p}_2) = \hat{p}_1 - \hat{p}_2 = 0.3942 - 0.6058 = -0.2115$$

$$\hat{V}(\hat{p}_1 - \hat{p}_2) = \hat{V}(\hat{p}_1) + \hat{V}(\hat{p}_2) - 2cov(\hat{p}_1, \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n} + 2\frac{\hat{p}_1\hat{p}_2}{n}$$

$$= 0.0023 + 0.0023 + 2(0.0023) = 0.0092$$

95% Confidence Interval $= (\hat{p}_1 - \hat{p}_2) \pm 2\sqrt{\hat{V}(\hat{p}_1) + \hat{V}(\hat{p}_2) - 2cov(\hat{p}_1, \hat{p}_2)} = -0.2115 \pm 2\sqrt{0.0092}$

$$= (-0.4032, -0.0199)$$

The 95% confidence interval of the difference does not include the value of zero.

Thus, there is a difference in proportion of participants who rate their health as satisfactory over non-satisfactory among the males at 5% significance level.

(vii) The prospective buyers that the company should approach in its future promotional activities:

The company should approach male customers, with the age between 23 to 25 years old. Their education level should have 16 years of schooling. Besides that, they should have partnered marital status and their fitness level is 3 (unsatisfactory level). The prospective buyers should have an income around RM40,000 to RM 60,000. The company should approach this type of customers in its future promotional activities.

Question 2.

(i) Calculate sample size:

$$\hat{p} = \frac{\sum_{i=1}^{4} a_i}{\sum_{i=1}^{4} m_i} = \frac{505}{1626} = 0.3106$$

$$s_p{}^2 = \frac{\sum_{i=1}^{4}(a_i - \hat{p}m_i)^2}{n-1} = \frac{1222.090976}{3} = 407.364$$

$$\bar{M} = \bar{m} = \frac{\sum_{i=1}^{4} m_i}{n} = \frac{1626}{4} = 406.5$$

$$D = \frac{B^2\bar{M}^2}{4} = \frac{(0.04^2)(406.5^2)}{4} = 66.097$$

$$n = \frac{Ns_p{}^2}{ND + s_p{}^2} = \frac{(18)(407.364)}{(18)(66.097)+407.364} = 4.59 \simeq 5$$

$\therefore$ Thus, 5 clinics should be selected.

(ii) The strengths and limitations of the sampling method proposed:

The sampling method proposed above is cluster sampling method.

Strengths:

Cluster sampling can help to save time and cost. Therefore it is more economical. Also, it can be used when no sampling frame is available. Besides that, it allows each accumulation of large samples. Furthermore, it can avoid overlapping, which means the case of redundancy will not happen.

Limitations:

Cluster sampling usually provides less precision. Sometimes it might have the difficulties to apply the findings to other areas. Besides that, an element of sample bias will arise when unequal size of the subsets is selected. In this question, since there are less clinics in the rural areas, when four clinics are randomly selected, rural areas will have a lesser chance to be selected. Hence, the sampling error will occur.

Question 3.

Let 1 represents A, 2 represents B, 3 represents C, 4 represents D.

$$\sum_{i=1}^{4} N_i \sqrt{\frac{p_i q_i}{c_i}} = 295.192 + 220.454 + 123.009 + 293.083 = 931.739$$

$$n_1 = \frac{295.192}{931.739} n = 0.317n$$

$$n_2 = \frac{220.454}{931.739} n = 0.237n$$

$$n_3 = \frac{123.009}{931.739} n = 0.132n$$

$$n_4 = \frac{293.083}{931.739} n = 0.315n$$

$$\sum_{i=1}^{4} N_i \sqrt{p_i q_i c_i} = 442.788 + 330.681 + 307.523 + 732.708 = 1813.701$$

$$\sum_{i=1}^{4} N_i p_i q_i = 129.094 + 108 + 74.723 + 181.794 = 493.611$$

$$D = \frac{B^2}{4} = \frac{0.01^2}{4} = 0.000025$$

$$n = \frac{\sum_{i=1}^{4} N_i \sqrt{\frac{p_i q_i}{c_i}} \times \sum_{i=1}^{4} N_i \sqrt{p_i q_i c_i}}{N^2 D + \sum_{i=1}^{4} N_i p_i q_i} = \frac{931.739 \times 1813.701}{(3375^2)(0.000025) + 493.611} = 2171.05 \simeq 2171$$

∴ The appropriate sample size is 2171.

The sample size for the four designs are listed below:

$n_1 = 0.317(2171) = 687.8 \simeq 688$ (Design A)

$n_2 = 0.237(2171) = 513.7 \simeq 514$ (Design B)

$n_3 = 0.132(2171) = 286.6 \simeq 287$ (Design C)

$n_4 = 0.315(2171) = 682.9 \simeq 683$ (Design D)

Question 4.

(i) Four perspectives of study:

1. Purpose of Survey
2. Development of Question Design
3. Development of Sampling Method
4. Development of Data Collection Method

The purpose of survey is to gather the information from Australian to examine the prevalence of ICD-10 disorders and associated comorbidity, disability and service utilization. Thus the respondents need to answer the survey accurately.

For the question design, the Composite International Diagnostic Interview (CIDI v2.1) is used in this study. Besides that, ICD-10 and DSM-IV are applied to identify disorders and cognitive impairment. The 12-item Short Form Health Survey (SF-12) and the National Comorbidity Survey days-out-of-role questions are also used to measure disability. Addition, UK Survey of Psychiatric Morbidity questions is applied to investigate perceived health. Therefore, the questions used in this study will perform well to collect the data.

Probability sampling is used in this case since the possible respondents are randomly selected. The sampling method used in this study is stratified multi-stage cluster sampling method. This will help obtain higher accuracy to represent the whole population.

The data collection is done through Composite International Diagnostic Interview and other measures. The interview is run through a laptop computer, which help reduce the anxiety of the participants when answering the questions. They will more likely to talk about their condition in this way. Besides that, there are screening questions to evaluate the existence of cognitive impairment and psychosis of the participants. If they are found having this kind of problems, they will not be included in the analysis. This is to ensure the accuracy of the responds. Furthermore, the data obtained is consistent with the previous parallel survey of low-prevalence disorderd. In addition, the interviewers are all well trained so that the answers recorded will be accurate.

Question 4.

(ii) The important aspects that should have been considered while developing questionnaire:

In this study, a screening questionnaire should be effective to filter unfavour respondents so that the analyses of the survey will be accurate. The next important thing is to make the questionnaire understandable. Since this is a health survey, it might include many professional terms. So, it should contain clear and understandable questions in the questionnaire. The structure of the questionnaire is also important. Also, it must be informative. The questionnaire should help obtaining the information needed. Furthermore, the questionnaire should be interesting so that the participants are willing to read through and answer it. Other than that, the questionnaire should be succinct. It should be straight to the point of the objective of the survey. Lastly, the questions designed in the questionnaire should be logical. More closed-ended questions is recommended to keep the responsed in a logic range.

(iii) The sampling method:

Stratified Multi-Stage Cluster Sampling Method.

The sampling and selection of participants:

The states and territories are the stratums, while the private dwellings are the clusters. There are two-stage sampling in this study. In the first stage, the private dwellings are randomly selected. In this study, 13624 private dwellings are randomly selected. Then in the second stage, one adult aged 18 years or above in each dwelling is being selected.

Since the group of Aborigines is undersampled, the age and gender characteristics of the sample are weighted to ensure the sample selected resembles the population. In addtion, the respondents need to answer the screening questions before being interviewed. If they are found to have for either cognitive impairment or psychosis through the screening questions, they are not allowed to participate the survey. This is to reduce the sampling error and. In addition, the survey excluded people in hospitals, nursing homes, hotels or jails, or residents of households in remote and sparsely settled parts of the country. This will help to ensure the sample selected resembles the population.

Question 4.

(iv) Points for the video presentation:

Sampling method:

Stratified Multi-Stage Cluster Sampling Method.

Strengths and Limitations:

It combines stratified sampling method and two-stage cluster sampling method. Thus, it generates higher accuracy as weighted sampling can be done. Besides that, it provides good coverage since each state and territory will have representatives. If the sampling frame for the dwellings is unavailable, the study still can be conducted since it uses cluster sampling in this case. Each dwelling has only one adult aged 18 years or above being selected. Therefore, it is also simple to be implemented. It will save time and also cost in this survey. Furthermore, the redundancy case, which means to measure the same person repeatedly, will not happen.

However, since multi-stage cluster is included in this study, sampling error might exist. Also, in second stage of cluster sampling, if the individual selected is unavailable (or nonresponse case), the household will not be represented, which means error occurs. Besides that, not all private dwellings will be selected in first stage of cluster sampling. If the types of dwellings is lesser, they will have a lower probability to be selected.

Steps to overcome the limitations:

To reduce the sampling errors in this study, we can seek ways to lower the the nonresponse cases. First, we can plan for callbacks in different time and different days to maximize the responses. On the other hand, we can provides rewards or incentives to encourage the responses. This is a very effective but expensive way. Furthermore, systematic sampling is the best way to obtain the high accuracy and precise results, but provided that the sampling frame is available, and also the time and funds are sufficient.

Video Submission Link: https://faudps2033.blogspot.com/