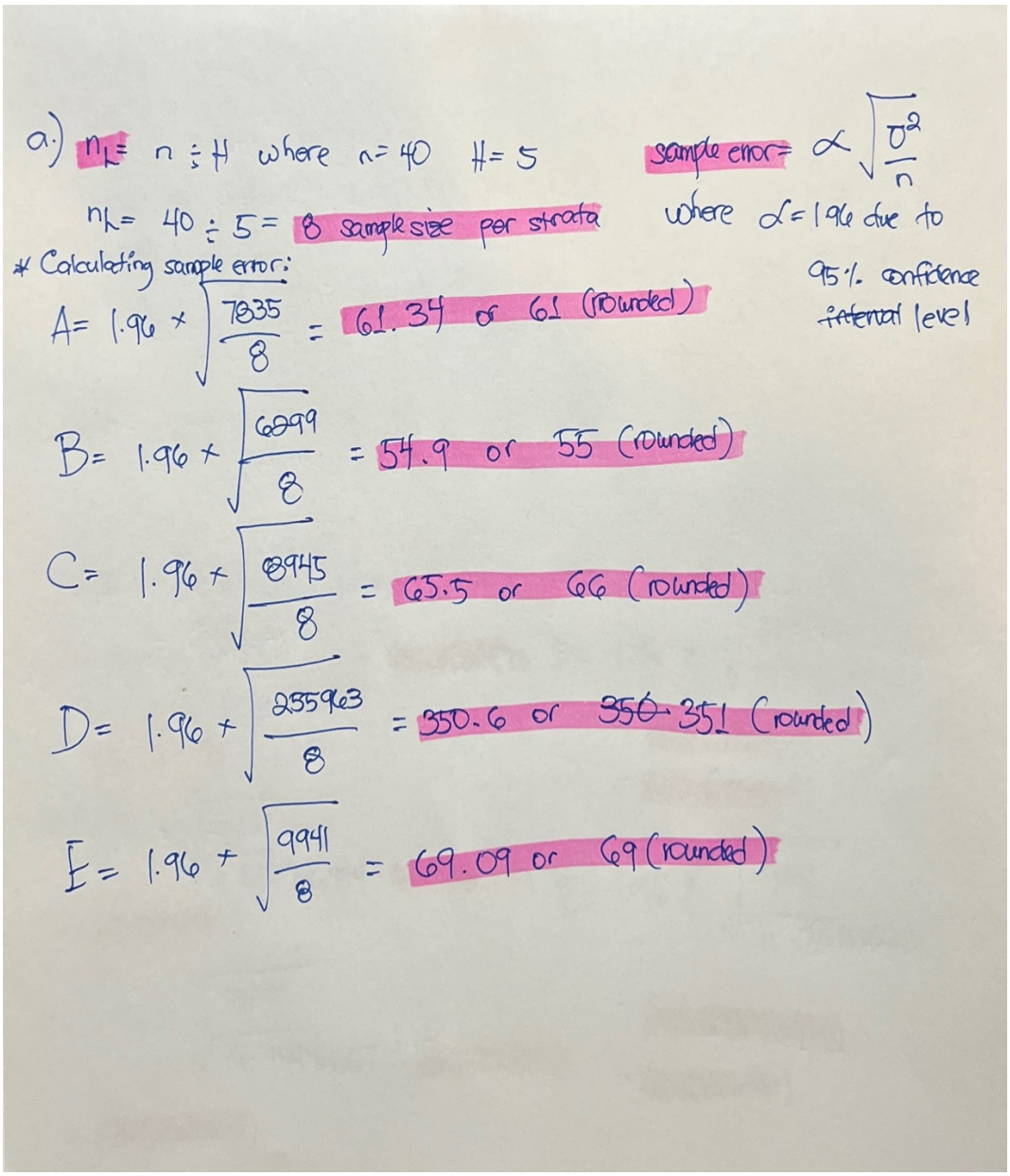
Question 1

a) **Calculate the expected sample error at a 95% confidence level if the number sampled per strata is the same.**



a.) $n_k = n \div H$ where $n = 40$ $H = 5$

$n_k = 40 \div 5 = 8$ sample size per strata

sample error $= \alpha \sqrt{\frac{v^2}{n}}$ where $\alpha = 1.96$ due to 95% confidence interval level

* Calculating sample error:

$A = 1.96 \times \sqrt{\frac{7835}{8}} = 61.34$ or 61 (rounded)

$B = 1.96 \times \sqrt{\frac{6999}{8}} = 54.9$ or 55 (rounded)

$C = 1.96 \times \sqrt{\frac{8945}{8}} = 65.5$ or 66 (rounded)

$D = 1.96 \times \sqrt{\frac{235963}{8}} = 350.6$ or 350 ~~350~~ 351 (rounded)

$E = 1.96 \times \sqrt{\frac{9941}{8}} = 69.09$ or 69 (rounded)

b) Use proportional allocation to calculate the number to be sampled in each strata. Calculate the sample error at the 95% confidence level.

b.) $n_h = n\left(\frac{N_h}{N}\right)$

* Calculating sample size for each strata

$A = \frac{155}{1401} \times 40 = 4.425410421$

$B = \frac{256}{1401} \times 40 = 7.309064954$

$C = \frac{599}{1401} \times 40 = 17.10206995$

$D = \frac{10}{1401} \times 40 = 0.2855103498$

$E = \frac{381}{1401} \times 40 = 10.87794433$

* Calculating sample error using their sample sizes for each strata

$A = 1.96 \times \sqrt{\frac{1885}{4.425410421}} = 82.5$ or 83 (rounded)

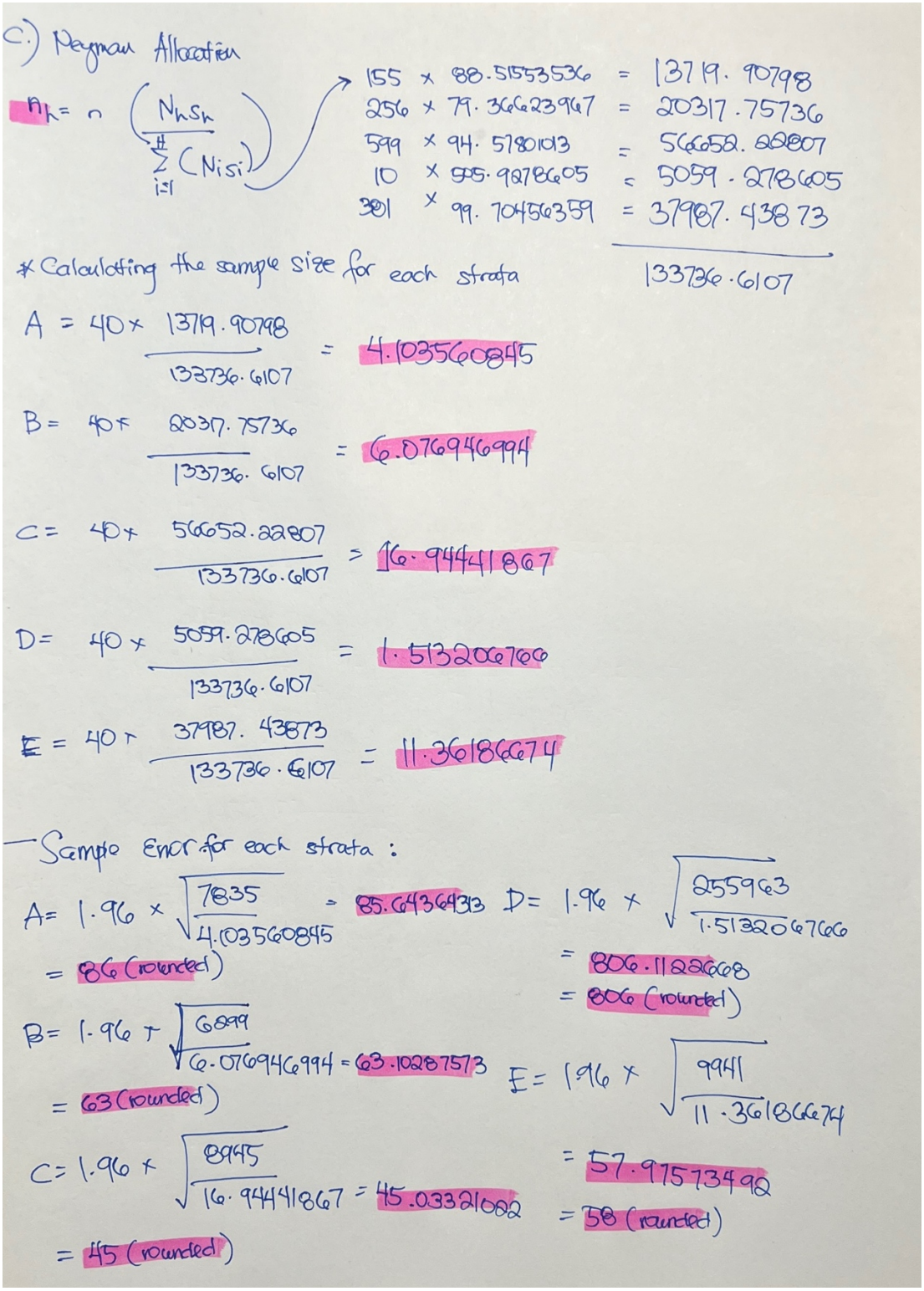
$B = 1.96 \times \sqrt{\frac{6899}{7.309064954}} = 57.5$ or 58 (rounded)

$C = 1.96 \times \sqrt{\frac{8945}{17.10206995}} = 44.8$ or 45 (rounded)

$D = 1.96 \times \sqrt{\frac{255963}{0.2855103498}} = 1855.8$ or 1856 (rounded)

$E = 1.96 \times \sqrt{\frac{9941}{10.87794433}} = 59.2$ or 59 (rounded)

c) Using Neyman allocation to calculate the number to be sampled in each strata. Calculate the sample error at the 95% confidence level.



c.) Neyman Allocation

$n_h = n\left(\frac{N_h S_h}{\sum_{i=1}^{H}(N_i s_i)}\right)$

155 × 88.51553536 = 13719.90798
256 × 79.36623967 = 20317.75736
599 × 94.5780103 = 56652.22807
10 × 505.9278605 = 5059.278605
301 × 99.70456359 = 37987.43873
133736.6107

* Calculating the sample size for each strata

$A = 40 \times \frac{13719.90798}{133736.6107} = 4.103560845$

$B = 40 \times \frac{20317.75736}{133736.6107} = 6.076946994$

$C = 40 \times \frac{56652.22807}{133736.6107} = 16.94441867$

$D = 40 \times \frac{5059.278605}{133736.6107} = 1.513206766$

$E = 40 \times \frac{37987.43873}{133736.6107} = 11.36186674$

— Sample Error for each strata :

$A = 1.96 \times \sqrt{\frac{7835}{4.103560845}} = 85.64364933$
= 86 (rounded)

$B = 1.96 \times \sqrt{\frac{6099}{6.076946994}} = 63.10287573$
= 63 (rounded)

$C = 1.96 \times \sqrt{\frac{8945}{16.94441867}} = 45.03321082$
= 45 (rounded)

$D = 1.96 \times \sqrt{\frac{255963}{1.513206766}}$
= 806.1122668
= 806 (rounded)

$E = 1.96 \times \sqrt{\frac{9941}{11.36186674}}$
= 57.97573492
= 58 (rounded)

d) Explain in less than ½ page why you think the results for the sample error in a), b) and c) are different.

In a), the sample sizes for each strata are the same as it assumes same variance for each strata. In b), the sample sizes for each strata are calculated with the assumption of different size strata population hence larger strata populations tend to get larger sample sizes and lower sample error. In c), the sample sizes for each strata are calculated with the assumption of different variances in strata hence larger strata variances tend to get larger sample sizes and lower sample error.

Different methods give different sample sizes for each strata hence different sample error for each strata. Usually, the sample error in c) is generally lower than b) and c) as it is more common to have different variances between each strata as opposed to different strata population (b) and same strata variance (a).

Question 2

a) Choose an SRSWOR sample of size 1,000 and then calculate the mean value for both variables along with their 95% confidence interval.

```
set.seed(31878039)
results = read.csv("Ass2Q2_data.csv")
var1 <- select(results, Popular_Culture_mark, math_mark)

# getting the simple random sample without replacement with 1000 sample size
s1=srswor(1000,125000)
s1_df <- as.data.frame(s1)
s1_df <- cbind(s1_df,var1)
s1_df_fin <- s1_df[s1_df$s1 == 1, ] # returning the data frame with the
selected
# 1000 sample sizes where rows in column s1 are equal to 1

# calculating math mark mean
est_pop_mean_math <- mean(s1_df_fin$math_mark)
est_pop_mean_math

## [1] 50.532

# calculating the 95% CI of math mark mean
s <- sd(s1_df_fin$math_mark)
n <- 1000
margin <- qt(0.975,df=n-1)*s/sqrt(n) #sample error
low_inter <- est_pop_mean_math - margin
low_inter

## [1] 49.15999

up_inter <- est_pop_mean_math + margin
up_inter
```

```
## [1] 51.90401

#calculating culture mark mean
est_pop_mean_culture <- mean(s1_df_fin$Popular_Culture_mark)
est_pop_mean_culture

## [1] 80.02832

# calculating the 95% CI of pop culture mark mean
s <- sd(s1_df_fin$Popular_Culture_mark)
n <- 1000
margin <- qt(0.975,df=n-1)*s/sqrt(n) # sample error
low_inter <- est_pop_mean_culture - margin
low_inter

## [1] 79.52775

up_inter <- est_pop_mean_culture + margin
up_inter

## [1] 80.52889
```

The mean for math marks is 50.5 with a confidence interval of between 49.2 and 52.

The mean for pop culture marks is 80 with a confidence interval of between 79.53 and 80.53.

b) Choose a single stage cluster sample of size 1,000 students and then calculate the mean value for both variables along with their 95% confidence interval.

```
set.seed(31878039)

# getting the 1000 sample sizes using simple random sample without
replacement
s1=srswor(1000,125000)
s1_df <- as.data.frame(s1)
s1_df <- cbind(s1_df,results)
s1_df_fin <- s1_df[s1_df$s1 == 1,] # returning the data frame with the
selected
# 1000 sample sizes where rows in column s1 are equal to 1

#setting up fpc = 5000 as that is the total number of cluster in the
population
s1_df_fin['fpc'] = 5000
# single stage cluster
api.onestage <- svydesign(id = ~cluster, data = s1_df_fin, fpc = ~fpc)
summary(api.onestage)

## 1 - level Cluster Sampling design
## With (911) clusters.
## svydesign(id = ~cluster, data = s1_df_fin, fpc = ~fpc)
```

```
## Probabilities:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1822  0.1822  0.1822  0.1822  0.1822  0.1822
## Population size (PSUs): 5000
## Data variables:
## [1] "s1"                  "math_mark"           "social_mark"
## [4] "cluster"             "student"             "Popular_Culture_mark"
## [7] "fpc"

# calculating the mean and 95% CI of math mark
summ_stats_math <- svymean(~math_mark, design = api.onestage)
summ_stats_math

##             mean     SE
## math_mark 50.532 0.6411

confint(summ_stats_math)

##             2.5 %  97.5 %
## math_mark 49.2755 51.7885

# calculating the mean and 95% CI of pop culture mark
summ_stats_culture <- svymean(~Popular_Culture_mark, design = api.onestage)
summ_stats_culture

##                         mean     SE
## Popular_Culture_mark 80.028 0.2309

confint(summ_stats_culture)

##                           2.5 %   97.5 %
## Popular_Culture_mark 79.57583 80.48082
```

The mean for math marks is 50.5 with a confidence interval of between 49.3 and 51.8.

The mean for pop culture marks is 80 with a confidence interval of between 79.6 and 80.5.

c) Choose a two-stage cluster sample of size 1,000 students with 5 students selected per cluster and then calculate the mean value for both variables along with their 95% confidence interval

```
set.seed(31878039)

s1=srswor(1000,125000)
s1_df <- as.data.frame(s1)
s1_df <- cbind(s1_df,results)
s1_df_fin <- s1_df[s1_df$s1 == 1,]

s1_df_fin['fpc'] = 5000
s1_df_fin['fpc2'] = 5
```

```
api.twostage <- svydesign(id = ~cluster + student, data = s1_df_fin, fpc =
~fpc + fpc2)
summary(api.twostage)

## 2 - level Cluster Sampling design
## With (911, 1000) clusters.
## svydesign(id = ~cluster + student, data = s1_df_fin, fpc = ~fpc +
##     fpc2)
## Probabilities:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03644 0.03644 0.03644 0.04322 0.03644 0.10932
## Population size (PSUs): 5000
## Data variables:
## [1] "s1"                   "math_mark"            "social_mark"
## [4] "cluster"              "student"              "Popular_Culture_mark"
## [7] "fpc"                  "fpc2"

#summ_stats_math <- svymean(~math_mark, design = api.twostage, na.rm = TRUE)
#summ_stats_math
#confint(summ_stats_math)

#summ_stats_culture <- svymean(~Popular_Culture_mark, design = api.twostage,
na.rm = TRUE)
#summ_stats_culture
#confint(summ_stats_culture)
```

d) Compare the results for the 3 different sampling processes across the 2 variables. Which would be the best and explain why you say this. You may want to consider the effort involved for each sampling scheme.

The results for a) and b) across the 2 variables for their mean along with their confidence interval are almost the identical. In this case, a) would be better to use as it is more straightforward and you don't need to use a new function (for the clusters) to give identical results. It is easier and takes less time as well to compute.