Task 3

Ariella Fuzaylov and Candice Djorno

4/3/2021

We transform the data Employee_B_probs into a matrix.

```
Employee.B.Branch=read.csv("Employee_B_by_Branch.csv", header=TRUE)
Employee.B.overall=read.csv("Employee_B_overall.csv", header=TRUE)
Employee.B.probs=as.matrix(read.csv("Employee_B_probs.csv", header=FALSE))
```

We add col and row names for human readability.

```
knitr::kable(head(Employee.B.Branch))
```

Year_Month	Branch	Total_of_Ratings	m_i
2011-12	Disneyland_California	4.469767	215
2011-12	Disneyland_HongKong	4.134831	89
2011-12	Disneyland_Paris	4.118421	76
2011-7	Disneyland_California	4.232877	73
2011-7	Disneyland_HongKong	4.083333	24
2011-7	Disneyland_Paris	3.791667	72

```
knitr::kable(head(Employee.B.overall))
```

Year_Month	Total_of_Ratings	m_i
2011-12	4.321053	380
2011-7	4.023669	169
2012-10	4.365155	419
2012-3	4.351351	370
2014-2	4.253968	252
2015-5	4.264012	678

```
month<-Employee.B.overall$Year_Month

colnames(Employee.B.probs)<-month
rownames(Employee.B.probs)<-month

knitr::kable(Employee.B.probs)
```

	2011-12	2011-7	2012-10	2012-3	2014-2	2015-5	2015-6	2015-7	2016-11	2016-2	2016-8	2017-11	2017-4	2018-12	2018-8
2011-12	0.0594317	0.0061980	0.0106966	0.0149327	0.0048871	0.0077698	0.0073959	0.0133081	0.0081639	0.0061392	0.0135883	0.0080456	0.0084400	0.0071992	0.0114919
2011-7	0.0061980	0.1121817	0.0202638	0.0282797	0.0092623	0.0147226	0.0140144	0.0252060	0.0154688	0.0116343	0.0257363	0.0152448	0.0159916	0.0136419	0.0217692
2012-10	0.0106966	0.0202638	0.1923618	0.0487390	0.0159819	0.0253951	0.0241746	0.0434500	0.0266810	0.0200718	0.0443626	0.0262951	0.0275820	0.0235327	0.0375335
2012-3	0.0149327	0.0282797	0.0487390	0.2669152	0.0223071	0.0354345	0.0337329	0.0605900	0.0372272	0.0280118	0.0618607	0.0366893	0.0384831	0.0328379	0.0523501
2014-2	0.0048871	0.0092623	0.0159819	0.0223071	0.0886201	0.0116104	0.0110518	0.0198815	0.0121990	0.0091744	0.0202999	0.0120224	0.0126115	0.0107580	0.0171696
2015-5	0.0077698	0.0147226	0.0253951	0.0354345	0.0116104	0.4403151	0.0175655	0.0315853	0.0193878	0.0145830	0.0322494	0.0191072	0.0200429	0.0170988	0.0272807
2015-6	0.0073959	0.0140144	0.0241746	0.0337329	0.0110518	0.0175655	0.1336334	0.0300681	0.0184556	0.0138815	0.0307004	0.0181884	0.0190792	0.0162765	0.0259698
2015-7	0.0133081	0.0252060	0.0434500	0.0605900	0.0198815	0.0315853	0.0300681	0.2384302	0.0331839	0.0249672	0.0551545	0.0327041	0.0343037	0.0292701	0.0466708
2016-11	0.0081639	0.0154688	0.0266810	0.0372272	0.0121990	0.0193878	0.0184556	0.0331839	0.1473484	0.0153221	0.0338815	0.0200753	0.0210583	0.0179653	0.0286619
2016-2	0.0061392	0.0116343	0.0200718	0.0280118	0.0091744	0.0145830	0.0138815	0.0249672	0.0153221	0.1111267	0.0254925	0.0151003	0.0158400	0.0135126	0.0215629
2016-8	0.0135883	0.0257363	0.0443626	0.0618607	0.0202999	0.0322494	0.0307004	0.0551545	0.0338815	0.0254925	0.2433535	0.0333917	0.0350248	0.0298856	0.0476508
2017-11	0.0080456	0.0152448	0.0262951	0.0366893	0.0120224	0.0191072	0.0181884	0.0327041	0.0200753	0.0151003	0.0333917	0.1452384	0.0207536	0.0177052	0.0282474
2017-4	0.0084400	0.0159916	0.0275820	0.0384831	0.0126115	0.0200429	0.0190792	0.0343037	0.0210583	0.0158400	0.0350248	0.0207536	0.1522718	0.0185724	0.0296295
2018-12	0.0071992	0.0136419	0.0235327	0.0328379	0.0107580	0.0170988	0.0162765	0.0292701	0.0179653	0.0135126	0.0298856	0.0177052	0.0185724	0.4301168	0.0252803
2018-8	0.0114919	0.0217692	0.0375335	0.0523501	0.0171696	0.0272807	0.0259698	0.0466708	0.0286619	0.0215629	0.0476508	0.0282474	0.0296295	0.0252803	0.3064285

Subtask 1

Estimate average rating

```
my_design<-svydesign(id=~Year_Month,prob=~diag(Employee.B.probs),
                     fpc=~rep(15/112,15),
                     data=Employee.B.overall,
                     pps=ppsmat(Employee.B.probs))
svymean(~Total_of_Ratings,my_design)
```

```
##                    mean     SE
## Total_of_Ratings 4.2182 0.0308
```

The estimated average satisfaction rating overall for the population of 40,041 reviews is 4.2182.

Confidence interval

```
conf= confint(svymean(x=~Total_of_Ratings,design = my_design))
conf
```

```
##                      2.5 %   97.5 %
## Total_of_Ratings 4.157762 4.278607
```

A 95% confidence interval is [4.1577624, 4.278607].

Subtask 2

Calculate Mean by Branch

```
knitr::kable(Employee.B.Branch%>%
  summarise(n= n(), Mean= mean(Total_of_Ratings),Var=sd(Total_of_Ratings)^2),caption = "Rating Summary 
```

Table 4: Rating Summary Statistics

n	Mean	Var
45	4.187407	0.0490165

```
knitr::kable(Employee.B.Branch%>% group_by(Branch)%>%
  summarise(n= n(), Mean= mean(Total_of_Ratings),StD=sd(Total_of_Ratings)), caption = "Rating Summarise
```

Table 5: Rating Summarised by Branch

Branch	n	Mean	StD
Disneyland_California	15	4.391302	0.1133633
Disneyland_HongKong	15	4.164182	0.1312240
Disneyland_Paris	15	4.006737	0.2094921

The estimated average rating for California is 4.391302, for HongKong is 4.164182, for Paris is 4.006737.

Hypothesis Test

We perform a hypothesis test to determine whether there is evidence that any of the ratings are statistically significantly different from each other in the population.

$H_0 : \mu_{california} = \mu_{hongkong} = \mu_{paris}$

$H_1 : \mu_{california} \neq \mu_{hongkong}$ or $\mu_{california} \neq \mu_{paris}$ or $\mu_{hongkong} \neq \mu_{paris}$ (i.e. the means are not all equal).

We perform an ANOVA.

```r
rating_aov = aov(Total_of_Ratings~Branch,data=Employee.B.Branch)
summary(rating_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Branch       2  1.121  0.5607   22.74 2.03e-07 ***
## Residuals   42  1.035  0.0247                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We obtain $p-value < 2.03e-07$ so $p-value < \alpha$. Therefore, we reject the null hypothesis and we conclude that there is evidence that Employee B could achieve more precision for these estimates.

Subtask 3

For Employee A: Overall estimated average rating: 4.2227 SE: 0.0125 95% confidence interval: [4.198217, 4.247116] California estimated average: 4.396533 HongKong estimated average: 4.213475 Paris estimated average: 3.976963 Result of ANOVA: the means are not all equal

For Employee B: Overall estimated average rating: 4.2182 SE: 0.0308 95% confidence interval: [4.157762, 4.278607] California estimated average: 4.391302 HongKong estimated average: 4.164182 Paris estimated average: 4.006737 Result of ANOVA: the means are not all equal

Let $\bar{y}_A$ be the estimated average for Employee A and $\bar{y}_B$ the estimated average for Employee B.

We observe that $SE(\bar{y}_A) < SE(\bar{y}_B)$ so $Var(\bar{y}_A) < Var(\bar{y}_B)$, therefore the estimate found by Employee A is more efficient than the estimate found by Employee B. Thus, the result found by Employee A provides the best answer. Because Employee B used months as clusters, this means people who went to the park in similar wheather would be in the same cluster. This would lead to homogeneity inside a single cluster, making cluster-sampling perform worse than SRSWOR.