Task 1

Ariella Fuzaylov and Candice Djorno

4/10/2021

```
Employee_A_data=read.csv("Employee_A_data.csv", header=TRUE)
```

Given

- Population of $N = 40,041$ reviews
- Employee A took an SRS of $n = 6,000$ reviews

Subtask 1:

Estimate average rating

Under SRSWOR, the sample mean $\bar{y} = \frac{1}{n}\sum_{i \in S} y_i$ is an unbiased estimator for the population mean $\hat{\mu}$.

```
N = 40041
n= 6000
y_bar<-sum(Employee_A_data$Rating)/n
```

Thus, the estimated average rating is $\hat{\mu} = 4.2226667$.

Confidence interval

```
srs_design = svydesign(id=~1,data=Employee_A_data, fpc=rep(N,n))
svymean(x=~Rating,design = srs_design)
```

```
##          mean     SE
## Rating 4.2227 0.0125
```

```
conf= confint(svymean(x=~Rating,design = srs_design))
conf
```

```
##           2.5 %   97.5 %
## Rating 4.198217 4.247116
```

A 95% confidence interval is [4.1982169, 4.2471165].

Subtask 2

Calculate Mean by Branch

```
Employee_A_data%>%
  summarise(n= n(), Mean= mean(Rating),Var=sd(Rating)^2)
```

```
##      n     Mean      Var
## 1 6000 4.222667 1.098269
```

```
Employee_A_data%>% group_by(Branch)%>%
  summarise(n= n(), Mean= mean(Rating),StD=sd(Rating))
```

```
## # A tibble: 3 x 4
##   Branch                    n  Mean   StD
##   <chr>                 <int> <dbl> <dbl>
## 1 Disneyland_California  2769  4.40 0.952
## 2 Disneyland_HongKong    1321  4.21 0.937
## 3 Disneyland_Paris       1910  3.98 1.19
```

```
# Employee_A_data%>% group_by(continent)%>%summarise(n= n(), Mean= mean(Rating),StD=sd(Rating))
```

The estimated average rating for California is 4.396533, for HongKong is 4.213475, for Paris is 3.976963.

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.

```
rating_aov = aov(Rating~Branch,data=Employee_A_data)
summary(rating_aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Branch         2    199   99.56   93.45 <2e-16 ***
## Residuals   5997   6389    1.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We obtain $p - value < 2e - 16$ so $p - value < \alpha$. Therefore, we reject the null hypothesis and we conclude that there is evidence that Employee A could achieve more precision for these estimates.