Task 2

Ariella Fuzaylov

4/11/2021

```
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:

```
Employee_A_data%>%
  summarise(n= n(), Mean= mean(Rating),Var=var(Rating))
```

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

```
strata.b<-Employee_A_data%>% group_by(Branch)%>%
  summarise(ni= n(), Mean= mean(Rating),Var=var(Rating))
strata.b
```

```
## # A tibble: 3 x 4
##   Branch                   ni  Mean   Var
##   <chr>                 <int> <dbl> <dbl>
## 1 Disneyland_California  2769  4.40 0.906
## 2 Disneyland_HongKong    1321  4.21 0.879
## 3 Disneyland_Paris       1910  3.98 1.43
```

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

```
## # A tibble: 5 x 4
##   continent    ni  Mean   Var
##   <chr>     <int> <dbl> <dbl>
## 1 Africa       66  4.15 1.33
## 2 Americas   2413  4.32 1.01
## 3 Asia        987  4.27 0.868
## 4 Europe     1772  4.04 1.33
## 5 Oceania     762  4.28 0.993
```

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.c = aov(Rating~continent,data=Employee_A_data)
summary(rating_aov.c)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## continent      4     88  21.894   20.19 <2e-16 ***
## Residuals   5995   6501   1.084
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

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

```
##               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
```

“Remember, stratification is most efficient when the stratum means differ widely; then the between sum of squares is large, and the variability within strata will be smaller. Consequently, when constructing strata we want the strata means to be as different as possible” pg 92 not wu

WE SHOULD FIX THIS

Reject both null hypotheses => means different

Stratification by Continent results in very similar sample means for each strata.

Chose stratification by Branch because this results in the largest difference between the sample means of each strata.

Subtask 2

Aim: optimally allocate sample sizes for a stratified sample of size 6,000.

Idea: use Neyman allocation with equal costs.

Proportional allocation assumes that the within stratum variance of a stratum is proportional to the size of the stratum. Meaning the larger the stratum the larger the within stratum variance. Therefore, to capture this variance accurately we take a larger sample from a larger stratum. From Employee A’s SRS of 6000 reviews we see that the mid size stratum has the highest variance, therefore we use Neyman allocation with allocates sample sizes proportional to the over all stratum times the stratum variance. This allocation will capture more of the variance in the sample.

```
n=6000
strata.b<-mutate(strata.b, Nh=c(19406,9619,13629))
denom=sum(strata.b$Nh*sqrt(strata.b$Var))
denom
```

```
## [1] 43762.16
```

```
numer=strata.b$Nh*sqrt(strata.b$Var)
nh=numer*n/denom
strata.b<-strata.b%>%mutate(nh=round(nh))
strata.b
```

```
## # A tibble: 3 x 6
##   Branch                   ni  Mean   Var    Nh    nh
##   <chr>                 <int> <dbl> <dbl> <dbl> <dbl>
## 1 Disneyland_California  2769  4.40 0.906 19406  2533
## 2 Disneyland_HongKong    1321  4.21 0.879  9619  1236
## 3 Disneyland_Paris       1910  3.98 1.43  13629  2231
```