Exercise 3

#Stratified Sampling

Note: The relative precision is given by

$$RP = \frac{V(\bar{y}_{srs}) - V(\bar{y}_{str})}{V(\bar{y}_{str})}(100)$$

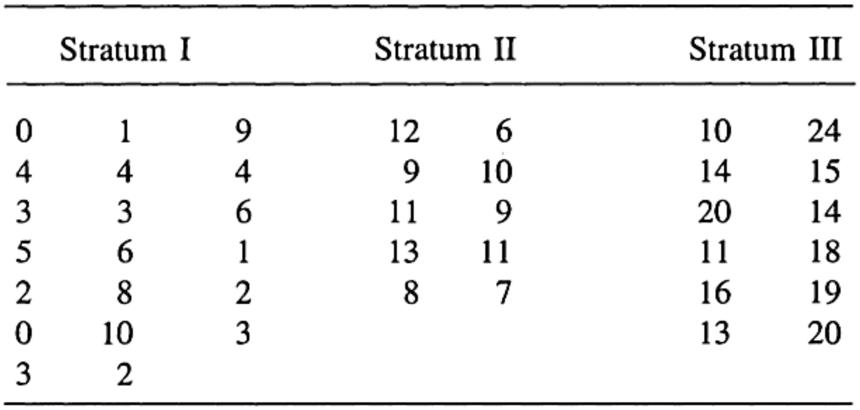
This gives an idea about the gain in efficiency due to stratification.

Example 5.1

An assignment was given to four students attending a sample survey course. The problem was to estimate the average time per week devoted to study in Punjab Agricultural University (PAU) library by the students of this university. The university is running undergraduate, master's degree and doctoral programs. Number of students registered for the three programs is 1300, 450, and 250 respectively. Since the value of the study variable is likely to differ considerably with the program, the investigator divided the population of students into 3 strata: undergraduate program (stratum I), master's program (stratum II), and doctoral program (stratum Ill). First of the four students selected WOR simple random samples of sizes 20, 10, and 12 students from strata I, II, and III respectively, so that, the total sample is of size 42. The information about weekly time devoted in library is given in table 5.1

1- Estimate the average time per week devoted to study by a student in PAU library. Also, build up the 95% confidence interval for this average.
2- Estimate the total time per week devoted to study by a students in PAU library. Also, build up the 95% confidence interval for this total.

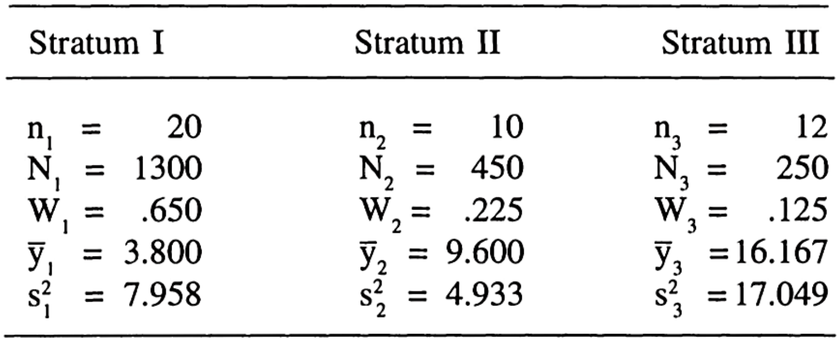
Table 5.1 Time (in hours) devoted to study in the university library during a week



Stratum I			Stratum II		Stratum III	
0	1	9	12	6	10	24
4	4	4	9	10	14	15
3	3	6	11	9	20	14
5	6	1	13	11	11	18
2	8	2	8	7	16	19
0	10	3			13	20
3	2					

Solution:

we first prepare table calculated values of strata sample means, weights and variance.



Stratum I	Stratum II	Stratum III
$n_1 = 20$	$n_2 = 10$	$n_3 = 12$
$N_1 = 1300$	$N_2 = 450$	$N_3 = 250$
$W_1 = .650$	$W_2 = .225$	$W_3 = .125$
$\bar{y}_1 = 3.800$	$\bar{y}_2 = 9.600$	$\bar{y}_3 = 16.167$
$s_1^2 = 7.958$	$s_2^2 = 4.933$	$s_3^2 = 17.049$

$$s_h^2 = \frac{\sum_{i=1}^{n}(y_{hi} - \bar{y}_{hi})^2}{n_h - 1} \;;\; \mathrm{W_h} = \frac{\mathrm{N_h}}{\mathrm{N}} \;;\; \mathrm{N} = \mathrm{N_1} + \mathrm{N_2} + \mathrm{N_3} = 2000$$

The estimate of average time (in hours) per week devoted to study by a student in the university library, is

$$\bar{y}_{str} = W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3 = 6.6508$$

Also, An unbiased Estimator of variance is

$$V\widehat{(\bar{y}_{str})} = v(\bar{y}_{str}) = \sum_{h=1}^{H} W_h^2 \frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

$$= 0.16553 + 0.02442 + 0.02113 = 0.21108$$

we obtain the limits of confidence interval as

$$\bar{y}_{str} \pm Z_{1-\propto/2}\sqrt{v(\bar{y}_{str})} \quad ;\; Z_{0.975} = 1.96$$

$$6.651 \pm 1.96\sqrt{0.21108}$$

$$[5.7505\ ,7.5515]$$

Note: Multiply these limits by N to obtain the limits for the population total (T).

We are 95% confidence that the average time per week devoted to study by a student in the PAU library belong in the closed interval $[5.7505\ ,7.5515]$

An unbiased estimator of the population total (total hours devoted to study by a students in PAU library)

$$\hat{t}_{str} = \sum_{h=1}^{H} N_h\bar{y}_h = N\,\bar{y}_{str} = 2000\,(6.6508) = 13301.6$$

Estimator of the variance of population total are obtained by

$$v(\hat{t}_{str}) = \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h} = N^2\, v(\bar{y}_{str})$$

$$v(\hat{t}_{str}) = (2000)^2(0.21108) = 844369$$

95%CI for total estimation

$$\hat{t}_{str} \quad \pm Z_{1-\propto/2} \sqrt{v(\hat{t}_{str})}$$

$$13301.6 \pm 1.96 \sqrt{844369}$$

$$[11501 \ \ , 15103]$$

By use R program

```
#Stratified Sampling
#open file (LBhour.txt)
LBhour<-read.delim("C:/Users/bayn/Desktop/LBhour.txt")
#please change the code here to your local address where you save the file
#hint: use "/" if "\" shows in the local address and does not work.
y<-LBhour$HOUR
stratum<-LBhour$PROGRAM
table(stratum)

stratum
 1  2  3 
20 10 12 

#compute sample mean and var for each stratum 1 ,2 and 3
#tapply:computes a measure or a function for each factor variable in a vector
tapply(y,stratum,mean)

       1        2        3 
 3.80000  9.60000 16.16667 

V=tapply(y,stratum,var)
V

        1         2         3 
 7.957895  4.933333 17.060606 

#Also, Following codes can help you look at mean and variance across stratum.
#sample size, mean and var for strata 1
y1<-y[stratum==1]
n1=length(y1)
y1_bar=mean(y1)
v1=var(y1)

#sample size, mean and var for strata 2
y2<-y[stratum==2]
n2=length(y2)
y2_bar=mean(y2)
v2=var(y2)

#sample size, mean and var for strata 3
y3<-y[stratum==3]
```

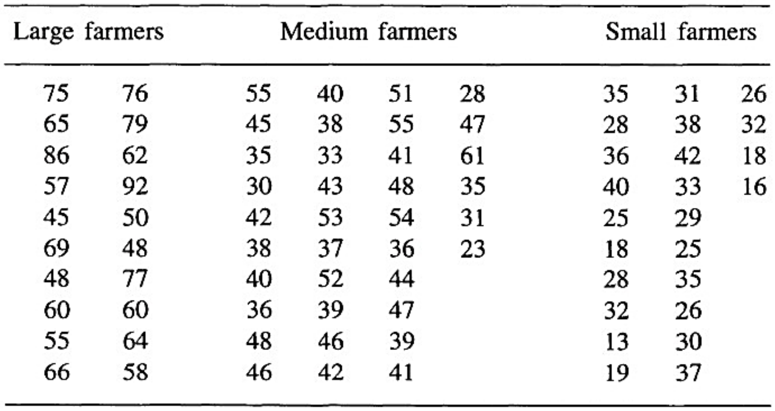
```
n3=length(y3)
y3_bar=mean(y3)
v3=var(y3)

N1=1300
N2=450
N3=250
N=N1+N2+N3
N

[1] 2000

#determine an estimate of the population mean by stratified sample
W1=N1/N; W2=N2/N ; W3=N3/N
str_mean=(W1*y1_bar+W2*y2_bar+W3*y3_bar)
str_mean

[1] 6.650833

##OR We can use following code
Wh=c(N1/N,N2/N,N3/N)
Wh[1]*y1_bar+ Wh[2]*y2_bar+ Wh[3]*y3_bar

[1] 6.650833

#Variance of sample mean =sum(((1-fh)/nh)*(Wh^2*Vh^2), fh=nh/Nh
var_str=((W1)^2)*((N1-n1)/N1)*(v1/n1)+((W2)^2)*((N2-n2)/N2)*(v2/n2)+((W3)^2)*
((N3-n3)/N3)*(v3/n3)
var_str

[1] 0.2110923

##OR We can use following code
#var_str=((W1)^2)*((N1-n1)/N1)*(V[1]/n1) +((W2)^2)*((N2-n2)/N2)*(V[2]/n2) +((
W3)^2)*((N3-n3)/N3)*(V[3]/n3)

#95%CI for mean  estimation
qnorm(1-(0.05/2)) #P(Z<a)=0.975>> a=1.96

[1] 1.959964

CI95_ybar=str_mean+(var_str^0.5)*(c(-1.96,1.96))
CI95_ybar

[1] 5.750316 7.551351

#estimation for total and its variance
t_hat=N*str_mean
t_hat

[1] 13301.67
```

```
Var_t=(N^2)*var_str
Var_t

[1] 844369

#95%CI for  total estimation
CI95_tot=t_hat+(Var_t^0.5)*(c(-1.96,1.96))
CI95_tot

[1] 11500.63 15102.70
```

Example 5.9

All the 80 farms in a population are stratified by farm size. The expenditure on the insecticides used during the last year by each farmer is presented in table 5.10 below:

Table 5.10 Expenditure (in '00 rupees) on insecticides used



Large farmers		Medium farmers				Small farmers		
75	76	55	40	51	28	35	31	26
65	79	45	38	55	47	28	38	32
86	62	35	33	41	61	36	42	18
57	92	30	43	48	35	40	33	16
45	50	42	53	54	31	25	29	
69	48	38	37	36	23	18	25	
48	77	40	52	44		28	35	
60	60	36	39	47		32	26	
55	64	48	46	39		13	30	
66	58	46	42	41		19	37	

a) Compute the overall population mean $\bar{Y}$ and the population mean square S^2.

b) Verify that population mean equal the weighted average of the means of the strata.

c) Determine the variance of the sample mean for stratified sampling with **proportional allocation** and **Equal allocation** of a sample of size n=24. Assume that the sampling is **WOR**.

d) Compare the relative precision of stratified sample mean $\bar{y}_{str}$ based on each of the above mentioned allocations, with respect to the simple random sample mean $\bar{y}_{srs}$.

e) Compare the relative precision of proportional allocation relative to Equal allocation.

f) Select a stratified sample of **24** farmers by using **Proportional allocation**.(use R program)

Solution: $N = 80$

stratum I (large farm)	stratum II (Medium farm)	stratum III (Small farm)
$N_1 = 20$	$N_2 = 36$	$N_3 = 24$
$\bar{Y}_1 = 64.6$	$\bar{Y}_2 = 42.1944$	$\bar{Y}_3 = 28.833$
$S_1^2 = 169.5158$	$S_2^2 = 70.5611$	$S_3^2 = 61.4493$
$w_1 = \frac{20}{80} = 0.25$	$w_2 = \frac{36}{80} = 0.45$	$w_2 = \frac{24}{80} = 0.30$

$$overall\ population\ mean:\ \bar{Y} = \frac{1}{80}(75 + 65 + \cdots + 16) = \ 43.7875$$

$$population\ mean\ square: S^2 = \frac{1}{80-1}[(75 - 43.7875)^2 + \cdots + (16 - 43.7875)^2]$$
$$= 268.6758$$

$$\bar{Y}_{str} = w_1\bar{Y}_1 + w_2\bar{Y}_2 \ + w_3\bar{Y}_3$$
$$\bar{Y}_{str} = 0.25(64.6) + 0.45(42.1944) + 0.30(28.8333) = 43.7875$$
$$\text{Therefore, } \ \bar{Y}_{str} = \bar{Y}$$

The sampling variance of mean in case of usual SRS without replacement is given by

$$V(\bar{y}_{srs}) = \frac{N-n}{Nn}S^2$$
$$= \frac{80-24}{80*24}(\mathbf{268.6758}) = 7.8364$$

We now obtain the value of $V(\bar{y})$ under two sample allocation methods:

Equal allocation: The number of units to be selected from each stratum will be $n_h = \frac{n}{H} = \frac{24}{3} = 8.$

The sampling variance for stratified mean

$$V(\bar{y}_{str}) = \sum_{h=1}^{H} w_h^2 \frac{S_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

$$= (0.25)^2\left(\frac{20-8}{20*8}\right)(169.5158) + (0.45)^2\left(\frac{36-8}{36*8}\right)(70.5611) + (0.30)\left(\frac{24-8}{24*8}\right)(61.4493)$$

$$= 2.6447$$

The relative precision of equal allocation based mean $\bar{y}_{str}$, with respect to usual mean $\bar{y}_{srs}$, is obtained as

$$PR = \frac{V(\bar{y}_{srs}) - V(\bar{y}_{str})}{V(\bar{y}_{str})}(100) = \frac{7.8364 - 2.6447}{2.6447} = 196\%$$

This result shows that there is a 196% gain in precision for equal allocation .

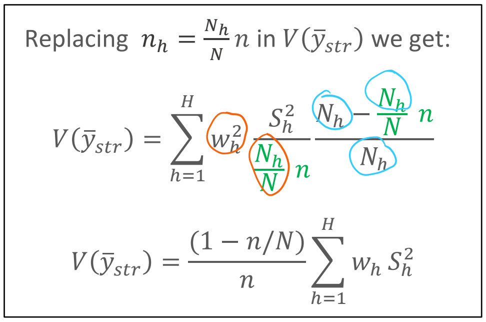
Proportional allocation. The number of units to be selected from each stratum will be

$$n_1 = \frac{N_1}{N}\ n = W_1\ n = 0.25 * 24 = 6$$

$$n_2 = W_2\ n = 0.45 * 24 = 10.8 \approx 11$$

$$n_3 = W_3\ n = 0.30 * 24 = 7.2 \ \approx 7$$

$$V(\bar{y}_{str.p}) = \frac{(1 - n/N)}{n}\sum_{h=1}^{H} w_h S_h^2$$



Replacing $n_h = \frac{N_h}{N}n$ in $V(\bar{y}_{str})$ we get:

$$V(\bar{y}_{str}) = \sum_{h=1}^{H} w_h^2 \frac{S_h^2}{\frac{N_h}{N}n}\frac{N_h - \frac{N_h}{N}n}{N_h}$$

$$V(\bar{y}_{str}) = \frac{(1 - n/N)}{n}\sum_{h=1}^{H} w_h\, S_h^2$$

$$V(\bar{y}_{str.p}) = \frac{1-24l80}{24}(0.25*169.5158+0.45*70.5611+0.30*61.4493) = 2.6998$$

The percent relative precision of proportional allocation based stratified mean $\bar{y}_{str}$ with respect to usual mean estimator $\bar{y}_{srs}$, is given by

$$PR = \frac{V(\bar{y}_{srs}) - V(\bar{y}_{str.p})}{V(\bar{y}_{str.p})}(100) = \frac{7.8364-2.6998}{2.6998}(100) = 190.26\%$$

This result shows that there is a 190% gain in precision for proportional allocation.

The relative precision of Proportional allocation relative to Equal allocation

$$PR = \frac{V(\bar{y}_{str}) - V(\bar{y}_{str.p})}{V(\bar{y}_{str.p})}(100) = \frac{2.6447-2.6998}{2.6998}(100) = |-2.04|\%$$

Thus, equal allocation reduces the variance by 2% from proportional allocation.

By use R program

```
#Stratified Sampling
#open file (farms.txt)
Farms<-read.delim("C:/Users/bayn/Desktop/KSU-STAT/farms.txt")
Y<-Farms$expenditure
stratum<-Farms$farms
table(stratum) #obtain size of each stratum.

stratum
 1  2  3 
20 36 24 

## compute population size, mean for each stratum 1,2 and 3.
#sample size, mean and var for strata 1
Y1<-Y[stratum==1]
N1=length(Y1)
Y1_bar=mean(Y1)
V1=var(Y1)
#sample size, mean and var for strata 2
Y2<-Y[stratum==2]
N2=length(Y2)
Y2_bar=mean(Y2)
V2=var(Y2)
#sample size, mean and var for strata 3
Y3<-Y[stratum==3]
N3=length(Y3)
Y3_bar=mean(Y3)
V3=var(Y3)
N=N1+N2+N3 #Population size
V1;   V2   ; V3
```

```
[1] 169.5158

[1] 70.56111

[1] 61.44928

## verify that population mean equal the weighted average of the means of the
strata
pop_mean=mean(c(Y1,Y2,Y3))  #or pop_mean=mean(Y)
pop_mean

[1] 43.7875

Wh=c(N1/N,N2/N,N3/N)
mean_str=Wh[1]*Y1_bar+Wh[2]*Y2_bar+Wh[3]*Y3_bar
mean_str

[1] 43.7875

# determine the variance of the sample mean for SRS with the same size 24,
Vsrs=(1-f)/n*S2, f=n/N

pop_v=var(c(Y1,Y2,Y3))  #or pop_v =var(Y)
pop_v
[1] 268.6758

n=24
Var_srs=((n-N)/(N*n))*pop_v
Var_srs

[1] 7.836377

#determine the variance of the sample mean for stratified sampling
with Equal allocation of a sample size n=24.
nh=24/3
nh

[1] 8

n1=n2=n3=nh
## v=sum((1-fh)/nh*(Wg^2*Vg^2))
Var_str.E=((Wh[1])^2)*((N1-n1)/N1)*(V1/n1)+((Wh[2])^2)*((N2-n2)/N2)*(V2/n2)+(
(Wh[3])^2)*((N3-n3)/N3)*(V3/n3)
Var_str.E

[1] 2.644647

#compare the precision of Equal allocation relative to simple random sampling
(Var_srs-Var_str.E)/Var_str.E *100

[1] 196.3109
```

```
#determine the variance of the sample mean for stratified sampling with
proportional allocation of a sample size n=24.


n1_p=round(Wh[1]*n, digits = 0) #Rounds a value to the nearest integer.
n2_p=round(Wh[2]*n, digits = 0)
n3_p=round(Wh[3]*n, digits = 0)
n1_p;  n2_p; n3_p

[1] 6

[1] 11

[1] 7

Sh2=c(var(Y1),var(Y2),var(Y3))
Var_str.P=(1-f)/n*sum(Wh*Sh2)
Var_str.P

[1] 2.699848

# compare the precision of proportional allocation relative to simple random
sampling.
(Var_srs-Var_str.P)/Var_str.P *100

[1] 190.2525

#Select a stratified SRS WOR of size 24 farmers by using Proportional
 allocation

set.seed(100)
sample1=sample(Y1,6)
sample1
[1] 66 69 48 92 79 48

sample2=sample(Y2,11)
sample2

 [1] 55 38 30 31 45 40 28 41 40 48 39

sample3=sample(Y3,7)
sample3

[1] 38 36 30 32 26 28 40

#determine an estimate of the population mean by stratified random sample .

str_mean=Wh[1]*mean(sample1)+Wh[2]*mean(sample2)+Wh[3]*mean(sample3)
str_mean

[1] 44.4026
```

# population proportion of Stratified Sampling

Example 5.12

The management of a local newspaper is to decide whether it should continue with the publication of 'Children Column', which had been introduced on experimental basis. For this purpose, it is imperative to estimate the proportion of readers who would favor its continuance. The frame consists of readers who had stayed with the paper for the last six months. The addresses of these readers are available in the office of the newspaper. Since different attitudes are expected from the urban and rural readers, it is reasonable to stratify the population into urban readers and rural readers. In the *population*, there are 73000 ***urban*** readers and 30280 ***rural*** readers. The investigator selected **WOR** simple random samples of 718 respondents from stratum **I** (urban readers) and 298 readers from stratum **II** (rural readers). The number of individuals who favor continuation of the column was 570 from stratum **I** and 143 from stratum **II.**

Estimate the proportion of readers interested in the continuation of the said column. Also, build up confidence interval for the population proportion.

Solution:

stratum I (urban readers)	stratum II (rural readers)
$N_1 = 73000$	$N_2 = 30280$
$n_1 = 718$	$n_2 = 298$
$x_1 = 570$	$x_2 = 143$
$w_1 = \frac{73000}{103280} = 0.7068$	$w_2 = \frac{30280}{103280} = 0.2932$
$p_1 = \frac{570}{718} = 0.7939$	$p_2 = \frac{143}{298} = 0.4799$

$N = N_1 + N_2 = 103280$

The estimate of the required overall population proportion is

$$\boldsymbol{p_{str} = \sum_{h=1}^{H} w_h p_h \quad = w_1 p_1 + w_2 p_2}$$

$$= (0.7068)(0.7939) + (0.2932)(0.4799) = 0.7018$$

The estimate of variance is

$$\boldsymbol{v(p_{str}) = \sum_{h=1}^{H} w_h^2 \frac{p_h q_h}{n_h} \frac{N_h - n_h}{N_h - 1}}$$

$$= w_1^2 \frac{p_1 q_1}{n_1} \frac{N_1 - n_1}{N_1 - 1} + w_2^2 \frac{p_2 q_2}{n_2} \frac{N_2 - n_2}{N_2 - 1}$$

$$= 0.0001127 + 0.00005129 = 0.0001839$$

The confidence interval for population proportion is

$$p_{str} \mp Z_{1-\frac{\propto}{2}} se(p_{str})$$

$$0.7018 \mp 1.96 \left(\sqrt{0.0001839}\right)$$

$$[0.67522\ ,\ 0.72837\ ]$$

To summarize, the sample estimate of proportion indicates that about 70 percent of reader population is in favor of continuing the column in question.

The population proportion favoring the continuance of the column will lie in the closed interval $[0.67522\ ,\ 0.72837\ ]$

Note:

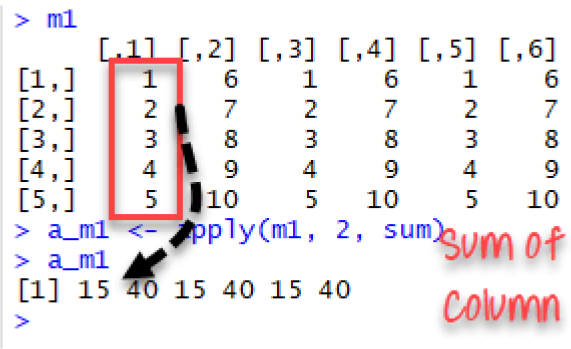
apply() and tapply() function in R

apply() takes Data frame or matrix as an input and gives output in vector, list or array.

```
apply(X, MARGIN, FUN)
Here:
-x: an array or matrix
-MARGIN: take a value or range between 1 and 2 to define where to apply the function:
-MARGIN=1: the manipulation is performed on rows
-MARGIN=2: the manipulation is performed on columns
-MARGIN=c(1,2): the manipulation is performed on rows and columns
-FUN: tells which function to apply. Built functions like mean, median, sum, min, max
and even user-defined functions can be applied.
```

Example:

```
m1 <- matrix(C<-(1:10),nrow=5, ncol=6)
a_m1 <- apply(m1, 2, sum)
a_m1
```



```
> m1
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    6    1    6    1    6
[2,]    2    7    2    7    2    7
[3,]    3    8    3    8    3    8
[4,]    4    9    4    9    4    9
[5,]    5   10    5   10    5   10
> a_m1 <- apply(m1, 2, sum)
> a_m1
[1] 15 40 15 40 15 40
>
```

Sum of Column

tapply() computes a measure (mean, median, min, max, etc..) or a function for each factor variable in a vector. It is a very useful function that lets you create a subset of a vector and then apply some functions to each of the subset.

```
tapply(X, INDEX, FUN = NULL)
Arguments:
-X: An object, usually a vector
-INDEX: A list containing factor
-FUN: Function applied to each element of x
```