A.Mayuri(2348133)

Analyses of New Workers in a Firm

A Study Using Real Life Data-set

Using(R-Markdown)

Analyses of New Workers in a Firm

A.Mayuri(2348133)

2023-09-05

# Problem Statement:

Take a real life data having variables as continuous and categorical. Consider it as population and draw a sample of appropriate size by applying the SRS technique and give the estimates for Population mean and population total for both continuous and categorical variables. Write a report on it. Report should include the following: Complete data description with proper definition of study characteristics. Formula of each estimates with procedure and Analysis using R programming. Complete interpretation. significance of study in real life

## Data Description:

We have a dataset of size 206 (Large sample) where it contains the information about the employees who have joined an MNC firm which has a branch in India during the last 6 months.

## Variable of Interest:

we have cgpa,age,hours of work,salary,expenditure are the quantitative data. while gender and additional exams taken are qualitative data

## Objective:

1. Understand the data holistically - Using summary function
2. use Descriptive analyses to express the data -Graphical representation
3. estimate the suitable sample size for the study
4. estimate population mean and population total for the available quantitative variables available.
5. Categorical Data analyses

## Definition:

Variables:

1)CGPA-GPA is calculated by using the number of grade points a student earns in a given period of time.

2)Expenditure-The act of spending money.

3)Additional Exam- Any exams that were taken on the self interest of the employee than the university requirement (Eg, CFA,FRM,CA,CSIS etc,)

## FORMULA:

Sample Mean :

A black symbols with a white background

Description automatically generated

Population Mean:

A black and white math equation

Description automatically generated with medium confidence

Standard error for SRSWOR

A math problem with square root and square root

Description automatically generated with medium confidence

Population Total estimation



# ANALYSIS

## Step1:

let us initialize and attach the dataset to the given R-Markdown.

### Code

library(readxl)  
Book <- read\_excel("C:/Users/mayur/Desktop/Mstat/tri sem 1/R/Book.xlsx")  
View(Book)  
attach(Book)

## Step2:

Descriptive data analyses: let us try to understand the given variables.

note: the summary gives a “character” as a remark when the variable is categorical in nature.

### Code

summary(Book)

## NAME age GENDER Hrs of work per day  
## Length:206 Min. :21.00 Length:206 Min. : 3.000   
## Class :character 1st Qu.:23.00 Class :character 1st Qu.: 3.000   
## Mode :character Median :24.00 Mode :character Median : 5.000   
## Mean :23.85 Mean : 6.689   
## 3rd Qu.:25.00 3rd Qu.:11.000   
## Max. :27.00 Max. :11.000   
## CGPA salary in lakhs wrote any additional exams  
## Min. : 5.000 Min. : 6.00 Length:206   
## 1st Qu.: 6.000 1st Qu.: 8.00 Class :character   
## Median : 8.000 Median : 9.00 Mode :character   
## Mean : 7.597 Mean : 9.68   
## 3rd Qu.: 9.000 3rd Qu.:12.00   
## Max. :10.000 Max. :14.00   
## expenditure per year in lakhs  
## Min. :3.20   
## 1st Qu.:3.20   
## Median :7.90   
## Mean :5.99   
## 3rd Qu.:8.00   
## Max. :8.80

#### Interpretation

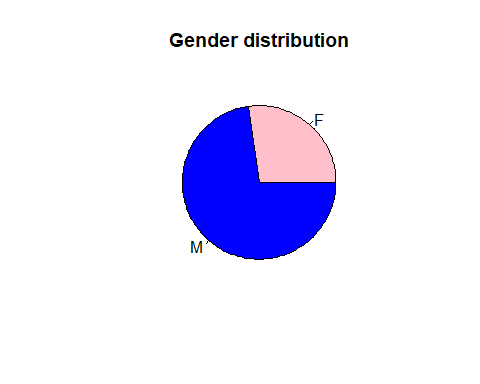
thus from the summary we have found the population mean for different continuous variables and the distribution of the data within different quantiles. Mean age= 23.85 Mean hours of work per day : 6.689 Mean CGPA : 7.597 Mean salary : 9.68 Mean expenditure :5.99

# Step2: Graph

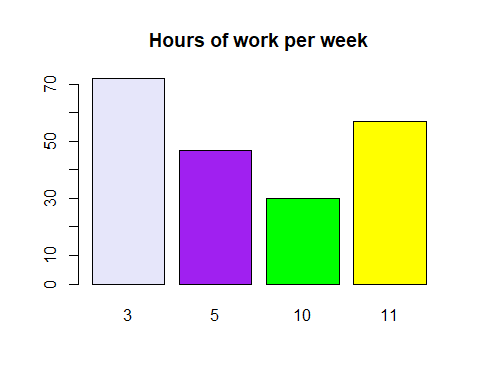
we can use different graphs to pictorially represent the data

### Code

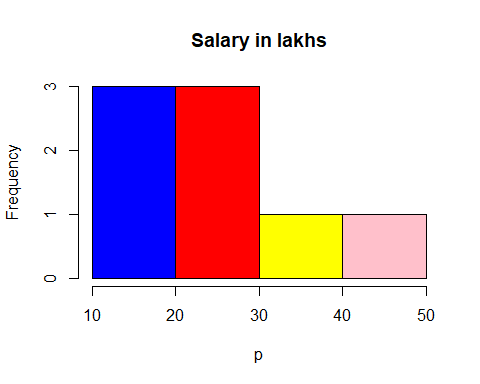
#Pie chart for gender  
m=table(Book$GENDER)  
pie(m,main='Gender distribution', col=c('pink','blue'))



#Bar chart for hours of work  
n=table(Book$`Hrs of work per day`)  
barplot(n, main="Hours of work per week",col=c("lavender","purple","green","yellow"))



# Histogram chart for salary information  
p=table(Book$`salary in lakhs`)  
hist(p,main="Salary in lakhs",col=c("blue","red","yellow","pink","purple","black","white","orange","lavender"))



## Step3: We have to calculate the sample size for the study

### Code

population=Book$`Hrs of work per day`  
N=length(Book$`Hrs of work per day`)  
N

## [1] 206

p=sd(population)  
p

## [1] 3.493629

S=N/(N-1)\*(p^2)  
S

## [1] 12.26499

sqrt(3.4936) #we have S as variance or S^2 in formula thus we tk sqrt of it to find d=z(a/2)\*SE

## [1] 1.869117

d=1.96\*1.869117  
d

## [1] 3.663469

n=((N\*(S^2))\*3.84)/((3.84\*S^2)+(N\*d^2))  
n

## [1] 35.60221

#### Interpretation:

Thus we need an approximate minimum of sample size of 35.60221~36

## Step4:

##### Hours of Work per day

Here we are trying to find if the sample mean is an unbiased estimator of the population, evaluvate the population total for the quantitaive data “hours of work per day”

### Code

pop\_mean=mean(Book$`Hrs of work per day`)  
pop\_mean

## [1] 6.68932

n=36  
set.seed(12)  
sampDist1 <- replicate(20000,mean(sample(Book$`Hrs of work per day`, n)))  
mean(sampDist1)

## [1] 6.688961

Sd\_srswor=sd(sampDist1)\*((N-n/(N\*n))^(1/2))  
Sd\_srswor

## [1] 7.649311

Pop\_total=mean(sampDist1)\*N  
Pop\_total

## [1] 1377.926

#### Interpretation:

Thus the sample mean is an unbiased estimator of the population mean where, sample mean=6.688961 population mean= 6.68932 ie, on an average the employees work about 6~7 hours a day and population total (Y) = 1377.926 is the total hours worked by all the employees worked together in a day.

##### CGPA

Here we are trying to find if the sample mean is an unbiased estimator of the population, evaluvate the population total for the quantitative data “CGPA”

### Code

pop\_mean2=mean(Book$CGPA)  
pop\_mean2

## [1] 7.597087

n=36  
set.seed(12)  
sampDist2 <- replicate(20000,mean(sample(Book$CGPA, n)))  
mean(sampDist2)

## [1] 7.597117

Sd\_srswor2=sd(sampDist2)\*((N-n/(N\*n))^(1/2))  
Sd\_srswor2

## [1] 3.090771

Pop\_total2=mean(sampDist2)\*N  
Pop\_total2

## [1] 1565.006

Thus the sample mean is an unbiased estimator of the population mean where, sample mean=7.597087 population mean= 7.597117 ie, on an average the employee GPA is about 7.6. with a standard deviation of about 3.090771

##### Salary in Lakhs

Here we are trying to find if the sample mean is an unbiased estimator of the population, evaluvate the population total for the quantitative data “Salary in Lakhs”

### Code

pop\_mean3=mean(Book$`salary in lakhs`)  
pop\_mean3

## [1] 9.679612

n=36  
set.seed(12)  
sampDist3 <- replicate(20000,mean(sample(Book$`salary in lakhs`, n)))  
mean(sampDist3)

## [1] 9.680476

Sd\_srswor3=sd(sampDist3)\*((N-n/(N\*n))^(1/2))  
Sd\_srswor3

## [1] 5.253572

Pop\_total3=mean(sampDist3)\*N  
Pop\_total3

## [1] 1994.178

Thus the sample mean is an unbiased estimator of the population mean where, sample mean=9.679612 population mean= 9.680476 ie, on an average the employee GPA is about 9.7 Lakhs or (9~10)Lakhs. with a standard deviation of about 5.253572 Or the population total is about 1994.178 Lakhs which could also be interpreted as the expenditure of the firm on human capital

##### Expenditure

Here we are trying to find if the sample mean is an unbiased estimator of the population, evaluvate the population total for the quantitative data “Expenditure of th eemployee in a year”

### Code

pop\_mean4=mean(Book$`expenditure per year in lakhs`)  
pop\_mean4

## [1] 5.989806

n=36  
set.seed(12)  
sampDist4 <- replicate(20000,mean(sample(Book$`expenditure per year in lakhs`, n)))  
mean(sampDist4)

## [1] 5.994517

Sd\_srswor4=sd(sampDist4)\*((N-n/(N\*n))^(1/2))  
Sd\_srswor4

## [1] 5.293196

Pop\_total4=mean(sampDist4)\*N  
Pop\_total4

## [1] 1234.871

sample mean= 5.989806 population mean= 5.994517 ie, on an average the employee GPA is about 5.9 Lakhs or ~(6)Lakhs. with a standard deviation of about 5.293196 or the population total is about 1234.871 lakhs per annum. or the employees spend about 1234.871 Lakhs per year.

## Categorical Data Analyses

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

df1= Book %>% group\_by(GENDER,`wrote any additional exams`) %>%  
 summarise(mean\_age=mean(age),mean\_Hrs\_work=mean(`Hrs of work per day`),mean\_cgpa =mean(CGPA), mean\_Salary=mean(`salary in lakhs`), mean\_expenditure=mean(`expenditure per year in lakhs`),.groups='drop') %>%  
 as.data.frame()  
df1

## GENDER wrote any additional exams mean\_age mean\_Hrs\_work mean\_cgpa  
## 1 F no 23.84615 7.564103 7.076923  
## 2 F yes 24.11765 7.941176 8.823529  
## 3 M no 22.98864 5.022727 7.340909  
## 4 M yes 25.00000 8.161290 7.951613  
## mean\_Salary mean\_expenditure  
## 1 8.435897 6.812821  
## 2 10.176471 6.782353  
## 3 8.397727 5.953409  
## 4 12.145161 5.306452

Conclusion: From the above table we can see the matrices that is generated by grouping the means of different quantitative paramenter with respect to each qualitative option(yes/no or F/M) ie, for eg, Exam vs No Additional Exam we have the mean age as to be 23.84 for females who did not take up any additional exams vs, the mean age of females who wrote exams as 24.11. This can be understood where students stay a little longer in education when they take up additional exams. also we can see that their average salary is considerably more (10.17 lakhs > 8.43 Lakhs) per annum for those who take up additional exams tend to earn more even in their first job.

Female Vs Male We significantly observe that a female is earning less despite taking up additional exams such that of a male conterpart. we observe that the difference may not be significant in the lower strata however as they climb the ladder of a firm the difference could be significant (male-12.14 Lakhs (vs) Female=10.17 Lakhs)

#### Significance of Study in Real life:

It is a true upholding that biased employment opportunities, salary etc are towards women, differently abled and similarly weaker section exsist. it is vital to understand and take measures to counterfeit it. Statistics as a science and tool for research helps us to validate the exsisting patterns, which can be addressed if collective measures are take.