

# PS1

## Applied Stats 1

Minh Trinh (Student ID: 24350478)

### Question 1

#### Section 1

Since number of observations = 25 which is smaller than 30, we will use a t-distribution with  $df = 24$  to calculate confidence interval. We calculate the sample mean, sampling error and t-value to construct the CI.

```
1 #Load data
2 y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98,
        80, 97, 95, 111, 114, 89, 95, 126, 98)
3 #Calculate lower and upper bound of confidence interval
4 sampling_mean <- mean(y)
5 sampling_sd <- sd(y)/sqrt(25)
6 t_value <- qt((1-0.9)/2, lower.tail = FALSE, df = 25 - 1)
7 CI <- c(sampling_mean - t_value*sampling_sd, sampling_mean + t_value*sampling_
        sd)
```

We have the result for 90% CI

93.95993 102.92007

#### Section 2

We follow the 5 step of hypothesis testing:

Step 1: Our assumption is that the data is randomly selected, the sample is relatively small ( $25 < 30$ ) and the data is quantitative. And according to the question, this will be an one-sided t-test

Step 2: Setting null and alternative hypothesis

$H_0$ : The school mean  $IQ \leq 100$

$H_a$ : The school mean  $IQ > 100$

Step 3: Calculating test statistics

```
1 sample_mean <- mean(y)
2
3 #calculate t-statistic
4 t = (sample_mean - 100)/(sd(y)/sqrt(25))
```

We have  $t = -0.596$

Step 4: Calculating p-value (in the direction of the alternative hypothesis)

```
1 #calculae p-val
2 pt(t, df = 24, lower.tail = FALSE)
```

We have p-value = 0.72 which is much larger than 0.05.

Step 5: Conclusion

We do not have enough evidence to reject the null hypothesis that the school student's mean  $IQ = 100$

## Question 2

Section 1

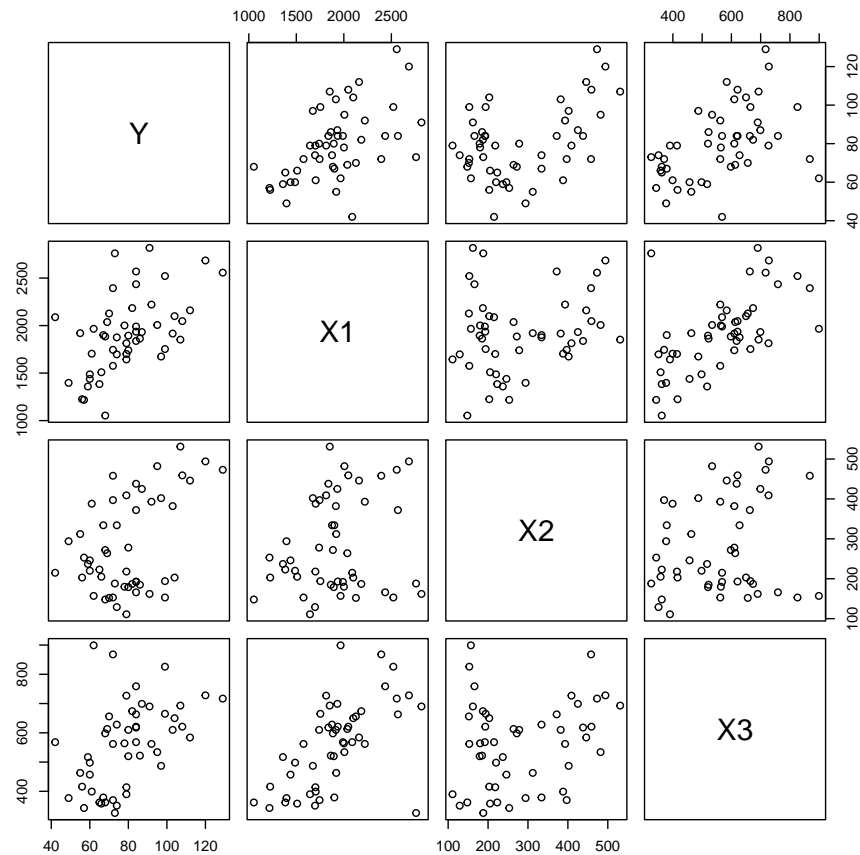
We create scatterplot pairwise and examine variables relationship

```
1 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
2 # scatter plot of X1,X2,X3,Y pairwise
3 pdf("/Users/tpminh/Desktop/trinity asds/stat analysis 1/ps1/pairplot.pdf")
4 pairs(expenditure[2:5])
5 dev.off()
```

As seen from Figures 1:

- X1 and Y seem to be positively correlated, as X1 increases, Y generally increases
- X2 and Y relationship changed as Y increase. When Y is at their lower values, as Y increase X2 decrease. When Y is at their higher values, as Y increase X2 increase
- X3 and Y seem to be positively correlated
- X1 and X2 do not seem to be correlated
- X1 and X3 seem to be positively correlated
- X2 and X3 do not seem to be correlated

Figure 1: Scatterplot between variables.



## Section 2

We use boxplot to examine relationship between Y and Region

```
1 pdf("/Users/tpminh/Desktop/trinity asds/stat analysis 1/ps1/boxplot.pdf")
2 boxplot(expenditure$Y ~ expenditure$Region)
3 dev.off()
```

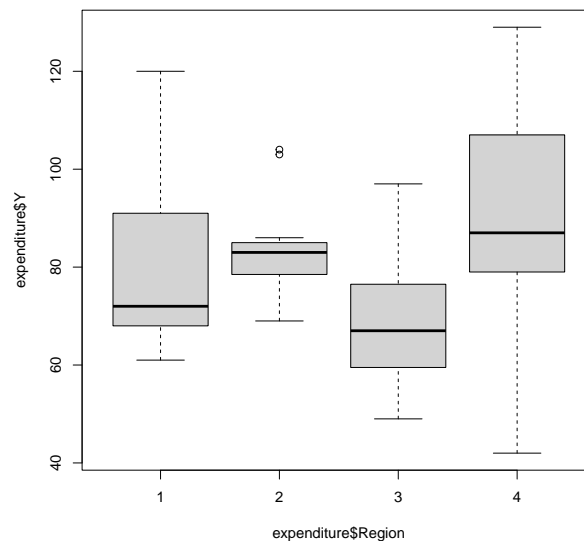
We can see that Region 4 has the highest median expenditures from Figures 2. But just to be sure, lets calculate the actual average of each region

```
1 aggregate(list(avg_exp = expenditure$Y), list(Region = expenditure$Region),
  FUN=mean)
```

Which give the result:

	Region	avg_exp
1	1	79.44444
2	2	83.91667
3	3	69.18750

Figure 2: Boxplot of expenditure by region.



4 4 88.30769

In conclusion Region 4 (West) have the highest expenditure of with the highest median and highest average of 88.3.

### Section 3

Lets examine relationship between Y and X1

```
1 plot( expenditure$X1,expenditure$Y)
```

As we have examined in the previous section, Y and X1 seem to be positively correlated (Figure 3). As X1 increase, Y will generally increase. Figure 4 changes the colour and symbol of scatterplot

```
1 plot( expenditure$X1,expenditure$Y, col = expenditure$Region, pch =
  expenditure$Region)
2 legend("topleft", c("1","2","3","4"), fill = 1:4)
```

Figure 3: Scatterplot between X1 and Y.

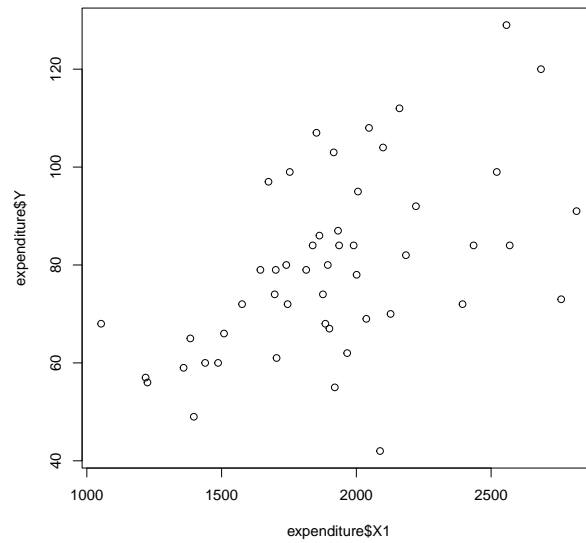


Figure 4: Scatterplot between X1 and Y with new colour and symbol.

