भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Factors Affecting Typing Speed

Project Presentation: MA4240 Applied Statistics

GROUP-5

April 25, 2022

Team Members

- Amulya Tallamraju - AI20BTECH11003
- Vaibhav Chhabra - AI20BTECH11022
- Anita Dash - MA20BTECH11001
- Anjali - MA20BTECH11002
- Ruthwika Boyapally - MA20BTECH11004
- Kunal Nema - MA20BTECH11007
- Sparsh Gupta - MA20BTECH11015
- Tapishi Kaur - MA20BTECH11017

Introduction

Abstract

Typing is now one of the essential part of almost every job since it lets you do your work very quickly and efficiently.
We aim to analyze the typing speed of students in IITH, try drawing conclusions based on various factors and analyze if there is some relation between the factors and the typing speed.

Collection of Data

Objective of the Study

Objective

The objective of our study is to ascertain factors affecting the typing speed of students.

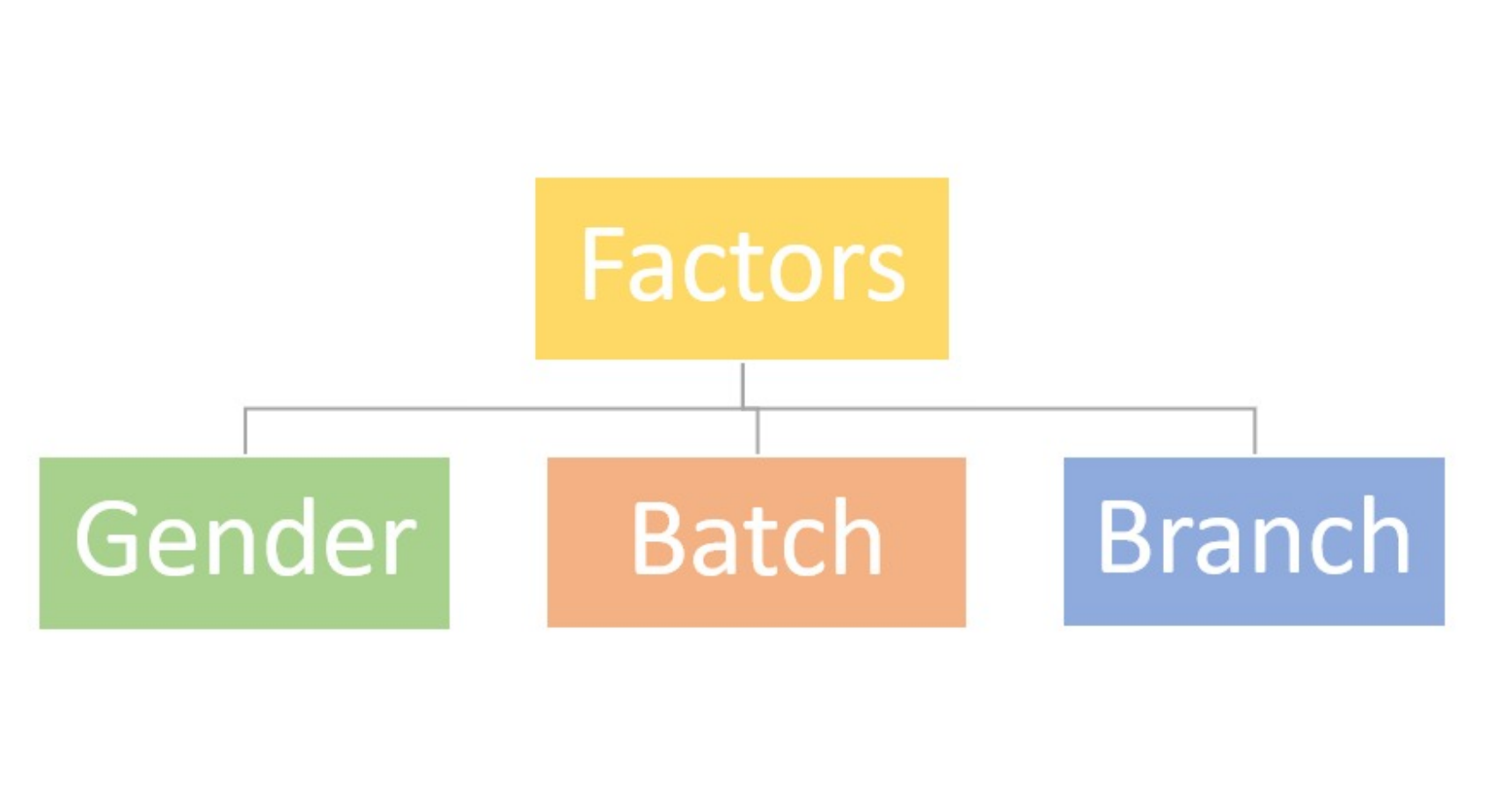


Figure 1: factors considered for the purpose of this study

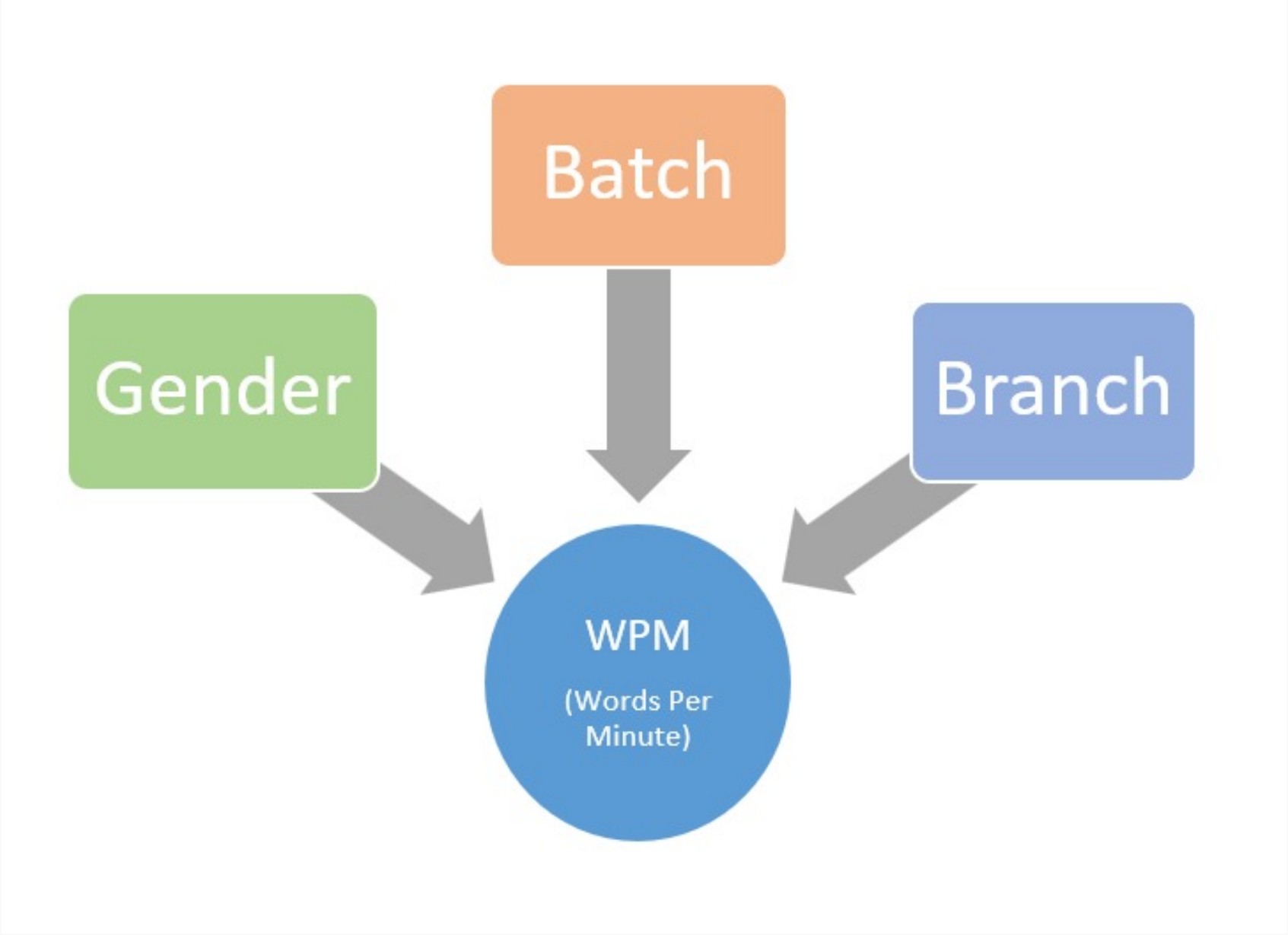


Figure 2: Is there a correlation between the factors and WPM?

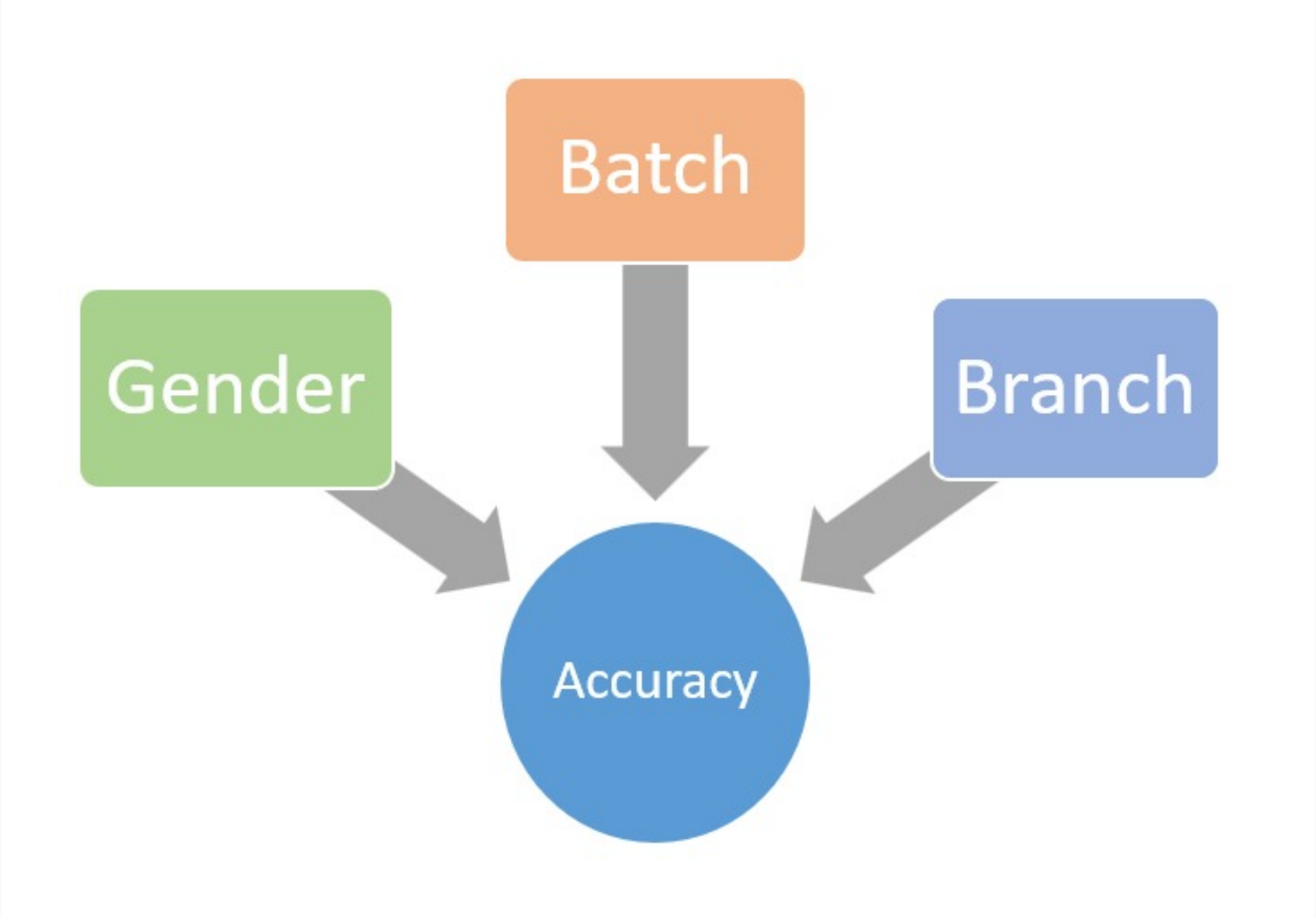


Figure 3: Is there a correlation between the factors and Accuracy?

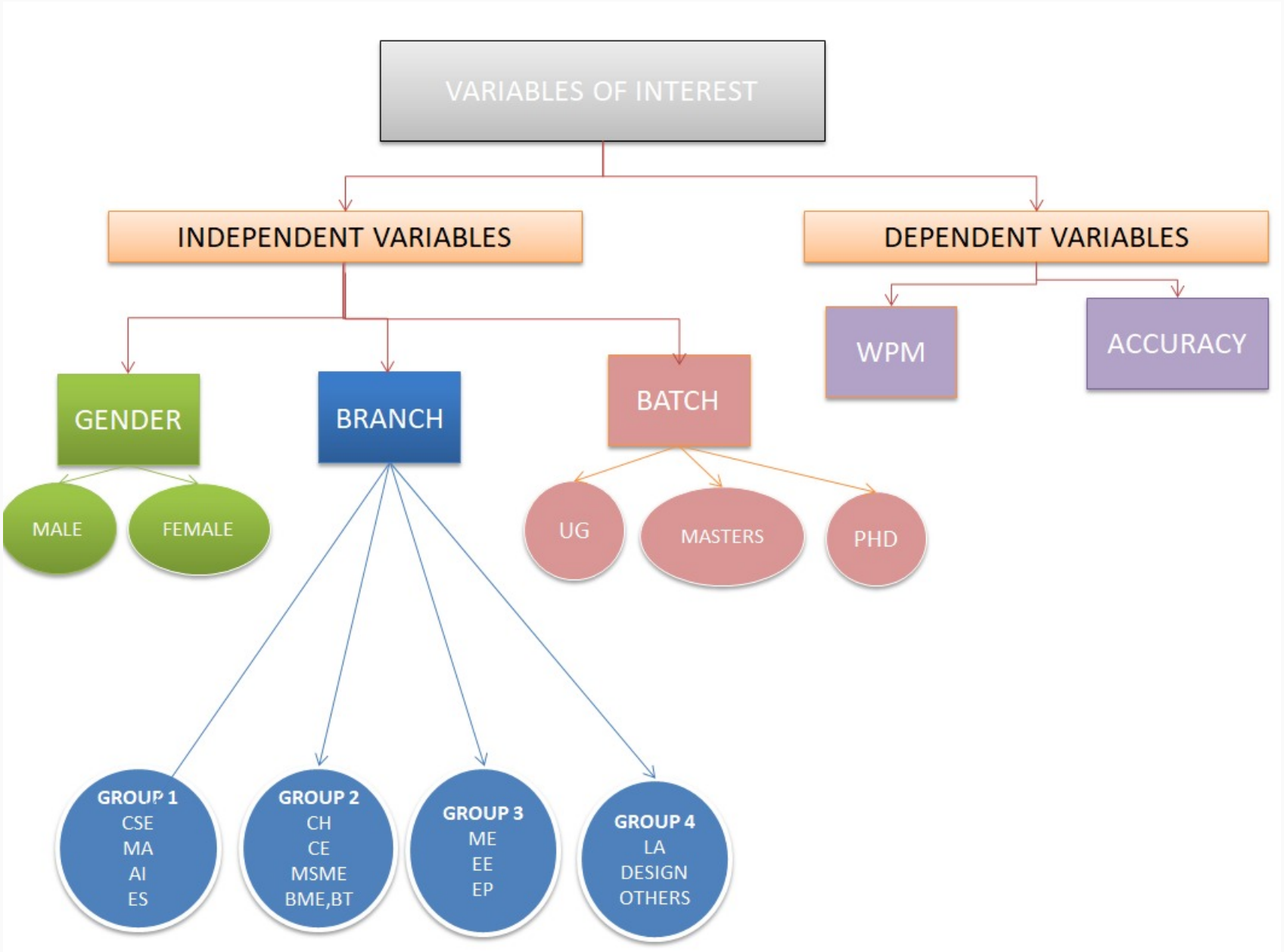


Figure 4: Variables of Interest

We perform an observational study to find out how the above stated factors affect the typing speed of an individual.

Strategy followed for Data Collection

- A google form was circulated to collect data.
- To avoid measurement problem in responses, participants were urged to take a common test.
- ▸ Link to the Test

Strategy followed for Data Collection

A Potential Issue

The survey we conducted is an example of convenience sampling. Hence, the data frame generated would be biased

Strategy followed for Data Collection

Definition

Simple random sampling is defined as a sampling technique where every item in the population has an even chance and likelihood of being selected in the sample.

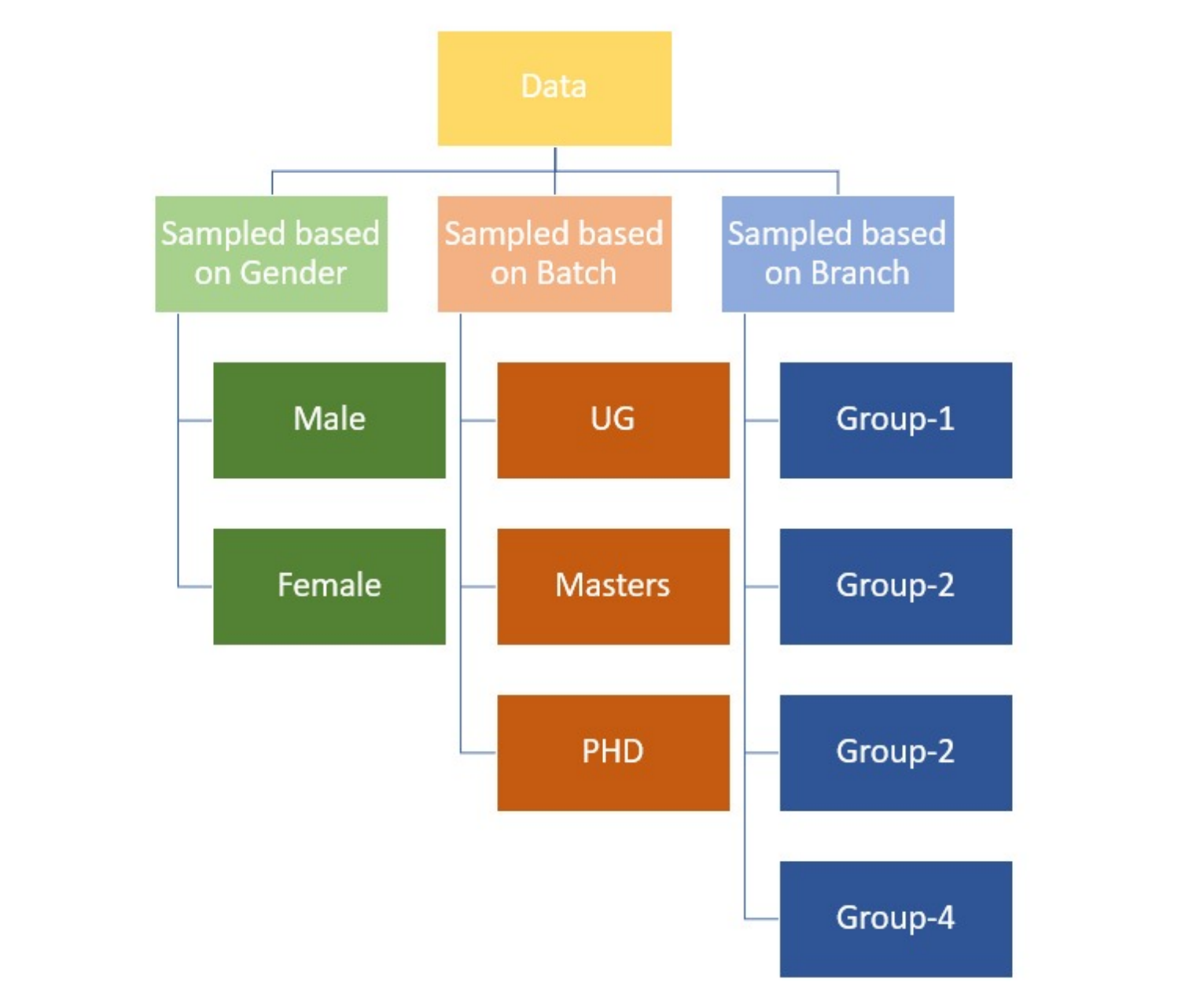


Figure 5: Simple random sampling of subgroups

Histogram Test

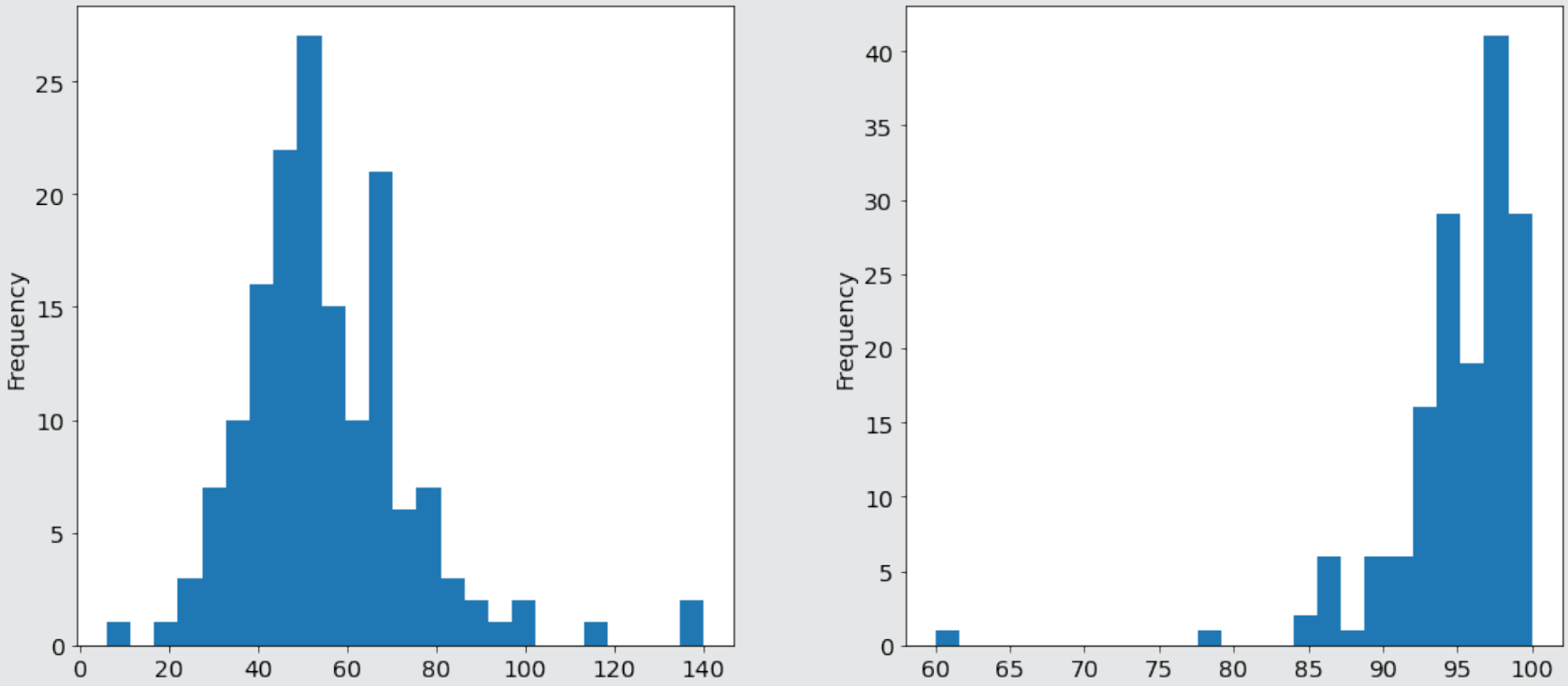


Figure 6: The graphs show that the data for wpm is right skewed and the accuracy is left skewed

Q-Q Plot Test

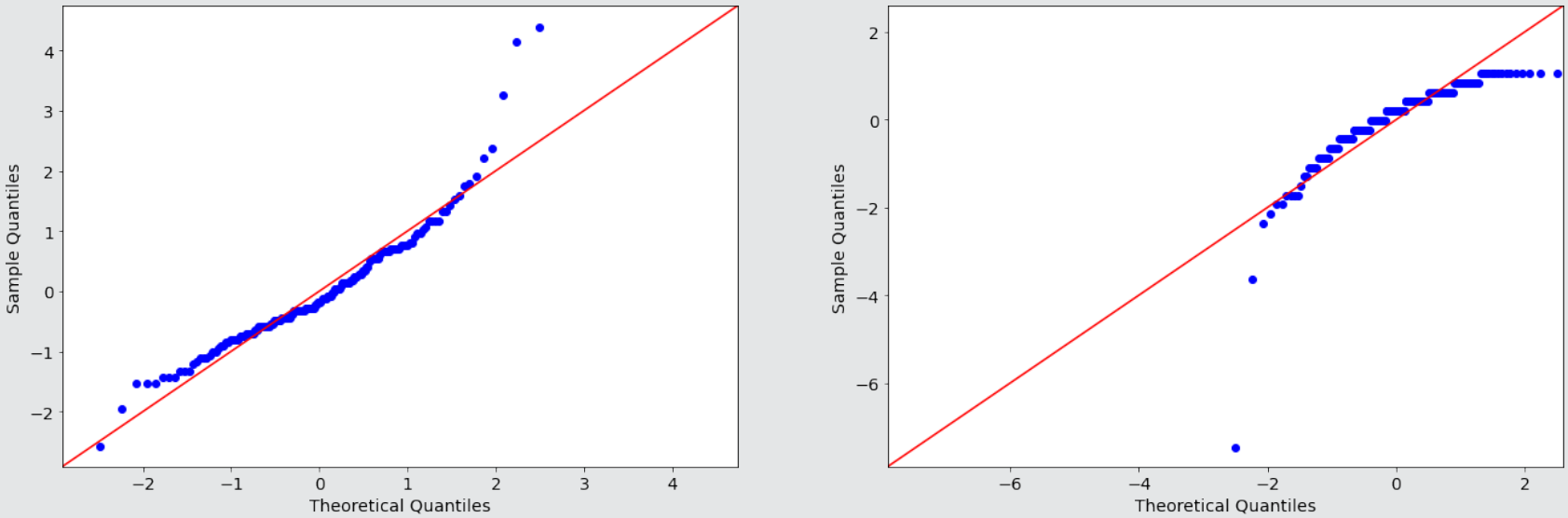


Figure 7: Certain number of points are away from the line

p-value test for normality

Hypothesis Testing with 0.05 level of significance More details

Null Hypothesis: Data is Normally Distributed

Alternative Hypothesis: Data is not Normally Distributed

```
[ ]  from scipy.stats import shapiro
     stat, p = shapiro(sample_data['WPM'])
     alpha = 0.05
     if p > alpha:
         print('Sample is normal (fail to reject H0)')
     else:
         print('Sample is not normal (reject H0)')
```

```
Sample is not normal (reject H0)
```

Issue

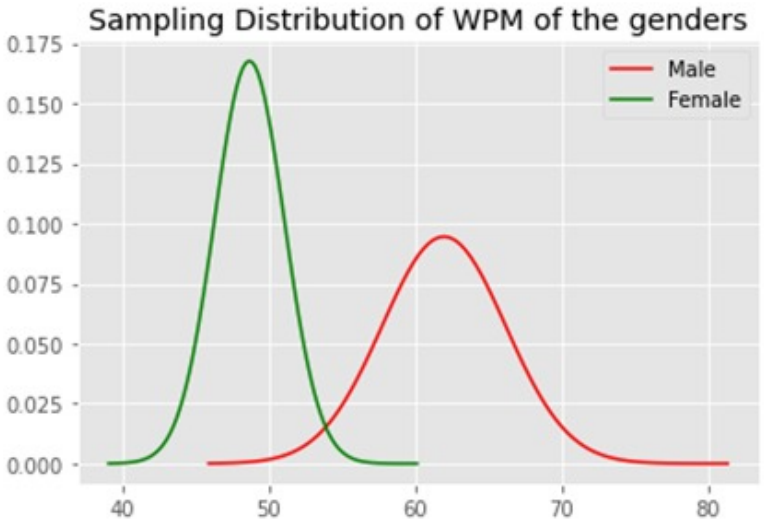
Our data is not normally distributed

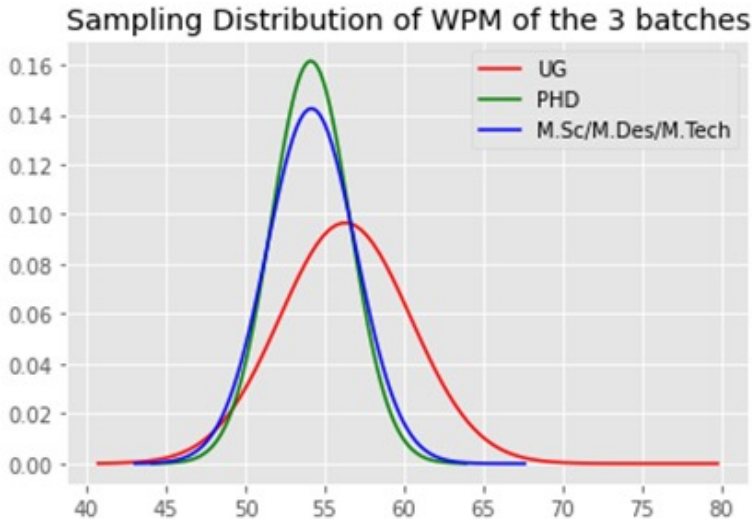
How do we go ahead with our study?

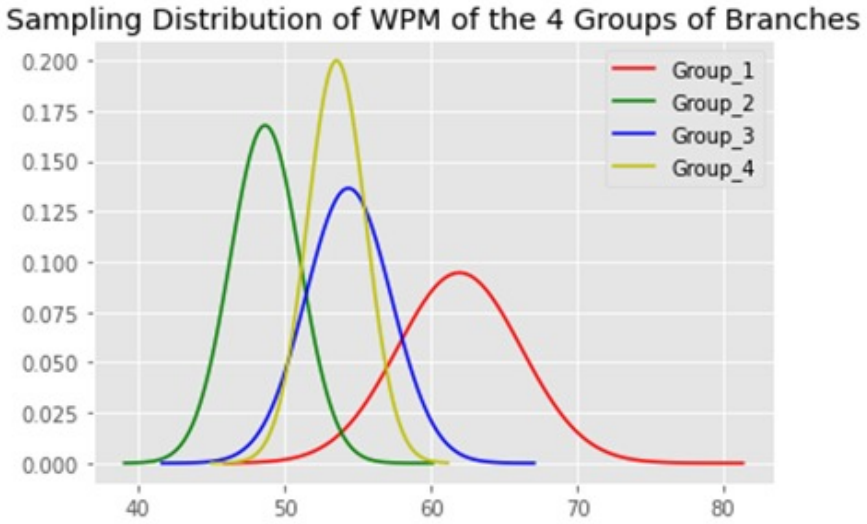
Central Limit Theorem

The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution. This fact holds especially true for sample sizes over 30.

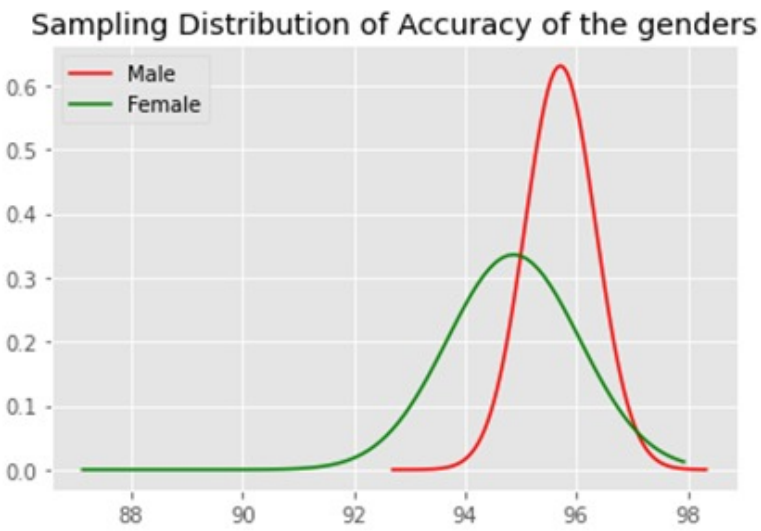
Sampling distribution for $n = 30$: WPM

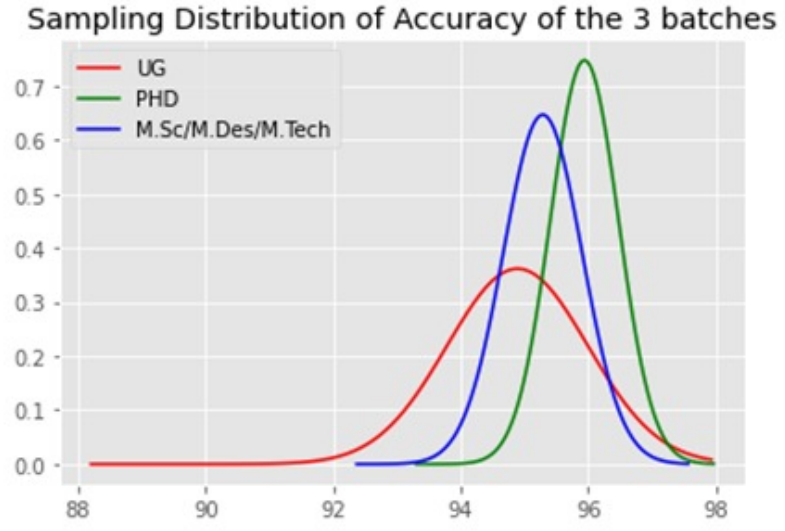


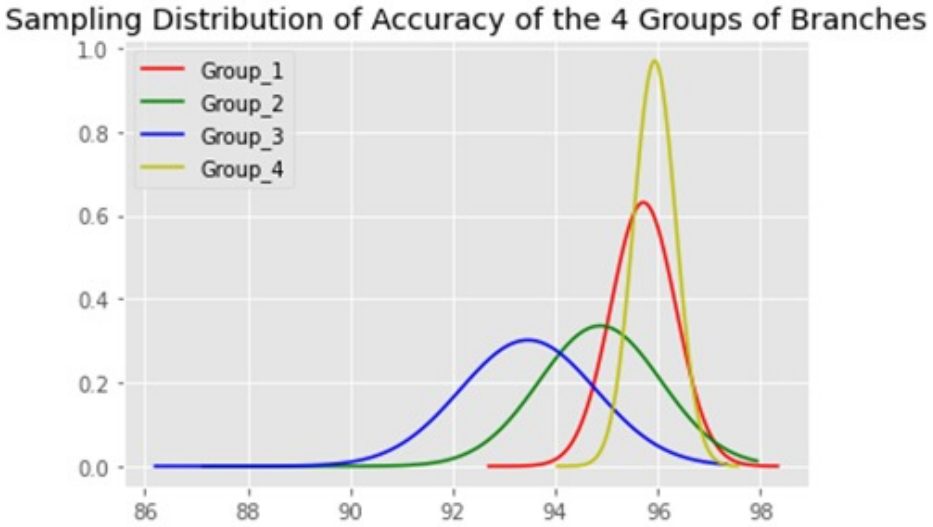




Sampling distribution for $n = 30$: Accuracy







Sample

	Gender	
	Male	Female
Total	107	50
Sampled	70	39

	Branch			
	Group-1	Group-2	Group-3	Group-4
Total	53	36	36	32
Sampled	45	30	30	30

	Batch		
	UG	M.Sc./M.Des/M.Tech	PHD
Total	80	41	36
Sampled	60	35	33

Figure 8: Data Collected and Data Sampled

Summarizing and Visualizing the Sample Data

Summarizing and Visualizing the Sample Data

Formulas

Sample Mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

Sample Variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (x_i - \bar{x})^2 \tag{2}$$

Gender-Based comparison

	Gender	
	Male	Female
Total	107	50
Sampled	70	39

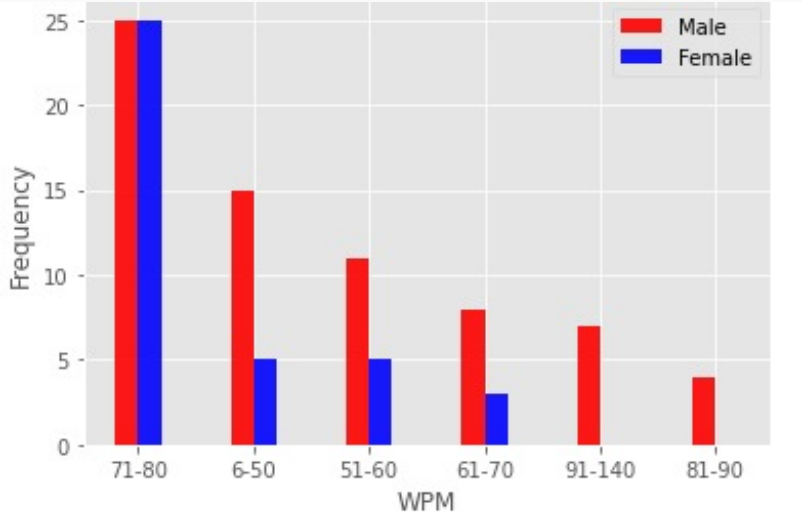
(a)

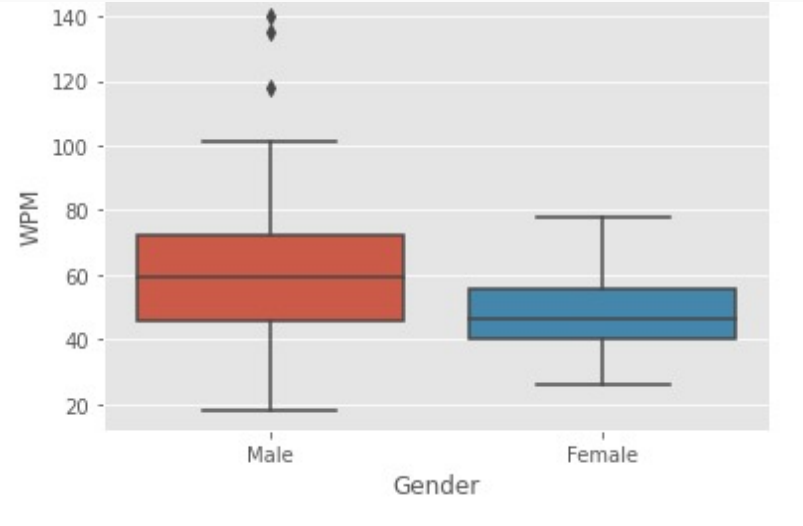
male	WPM	Accuracy
mean	61.985714	95.714286
variance	539.231677	12.236025

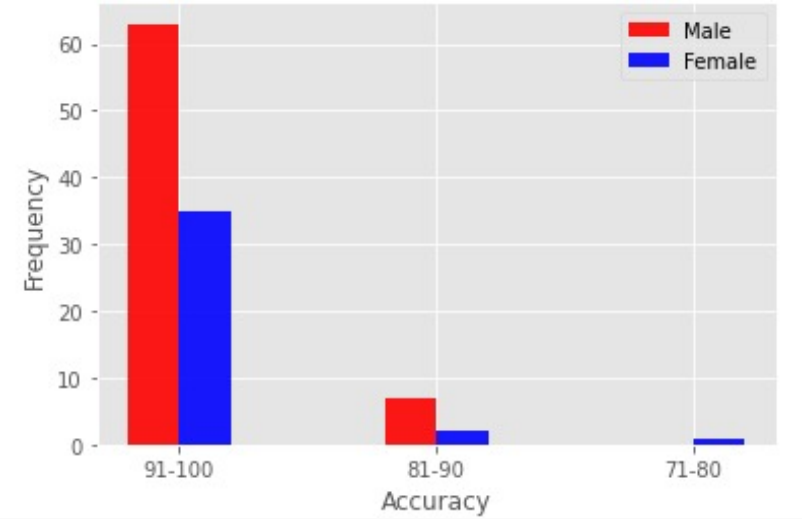
(b)

female	WPM	Accuracy
mean	49.179487	94.868421
variance	180.203779	43.901138

(c)


Male
Female
Frequency
WPM
71-80
6-50
51-60
61-70
91-140
81-90


WPM
Male
Female
Gender


Male
Female
Frequency
91-100
81-90
71-80
Accuracy

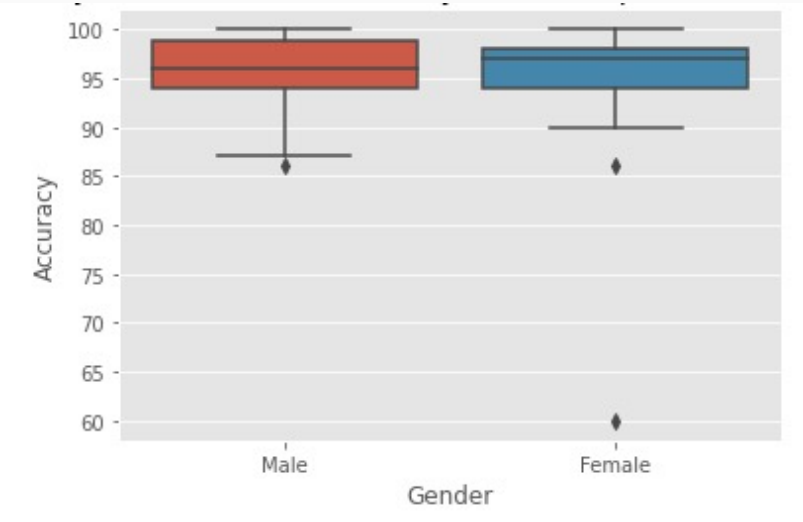

Accuracy
Male
Female
Gender

Figure 10

Batch-Based comparison

	Sample	
UG	**WPM**	**Accuracy**
mean	56.316667	94.883333
PHD	**WPM**	**Accuracy**
mean	54.121212	95.939394
Masters	**WPM**	**Accuracy**
mean	54.171429	95.285714

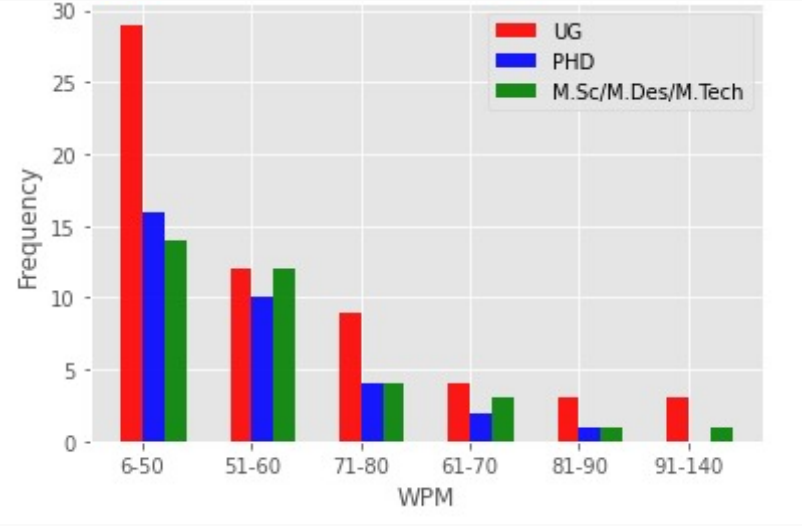
(a)

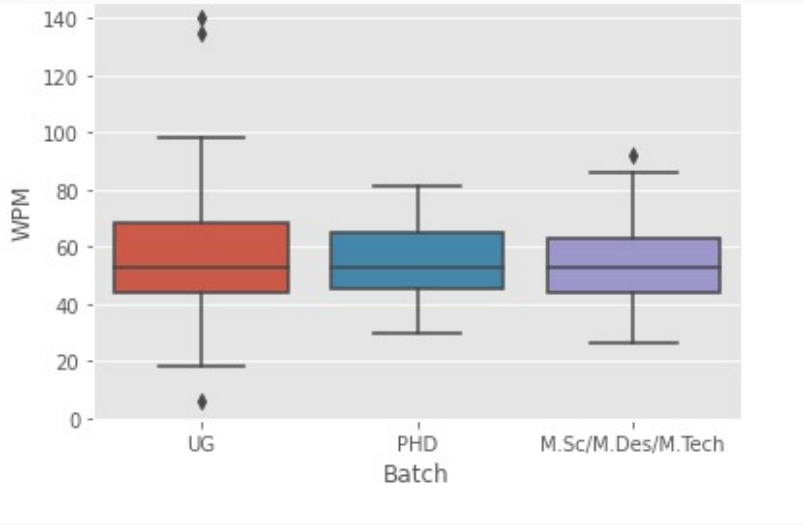
	Batch		
	UG	M.Sc./M.Des/M.Tech	PHD
Total	80	41	36
Sampled	60	35	33

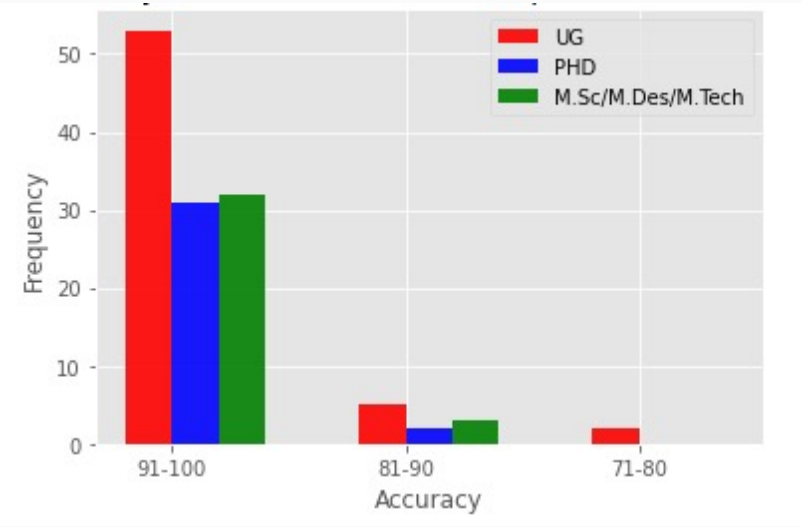
(b)

	Sample	
UG	**WPM**	**Accuracy**
variance	525.169209	37.392938
PHD	**WPM**	**Accuracy**
variance	187.672348	8.871212
Masters	**WPM**	**Accuracy**
variance	242.49916	11.857143

(c)


UG
PHD
M.Sc/M.Des/M.Tech
Frequency
WPM
6-50
51-60
71-80
61-70
81-90
91-140


WPM
Batch
UG
PHD
M.Sc/M.Des/M.Tech


UG
PHD
M.Sc/M.Des/M.Tech
Frequency
Accuracy
91-100
81-90
71-80

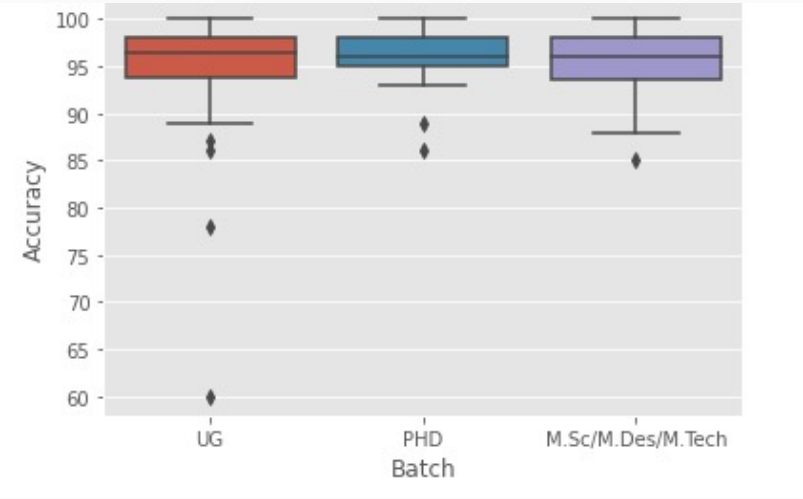

Accuracy
Batch
UG
PHD
M.Sc/M.Des/M.Tech

Figure 12

Branch-Based comparison

WPM	
Group	mean
G-1	66.29
G-2	45.33
G-3	54.37
G-4	53.57

(a)

	Branch			
	Group-1	Group-2	Group-3	Group-4
Total	53	36	36	32
Sampled	45	30	30	30

(b)

WPM	
Group	variance
G-1	650.66
G-2	120.64
G-3	262.52
G-4	123.43

(c)

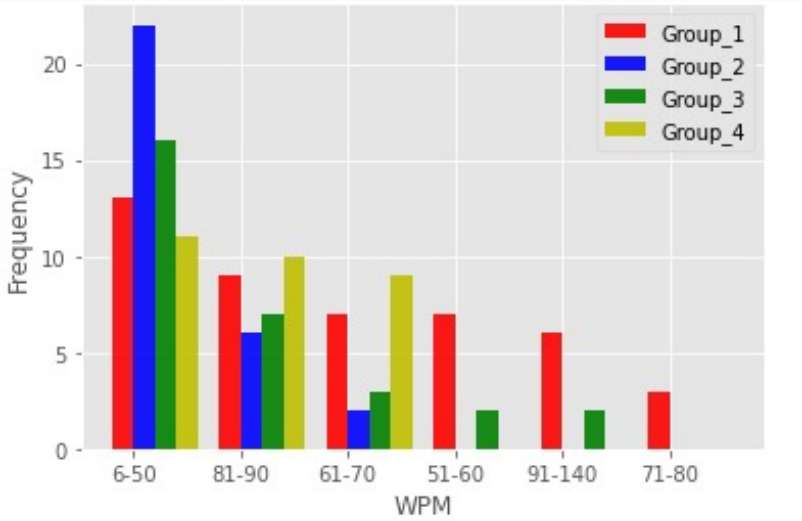
Branch-Based Comparision

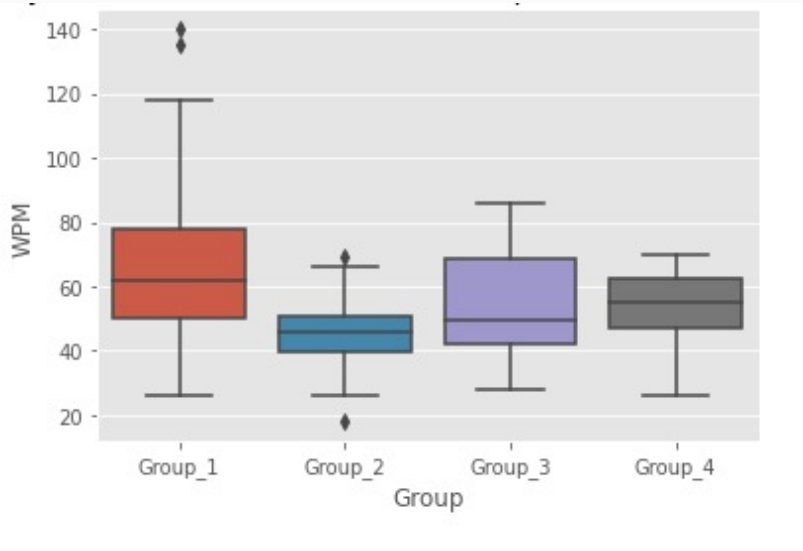
Accuracy	
Group	variance
G-1	11.904
G-2	17.66091954
G-3	53.91
G-4	5.24

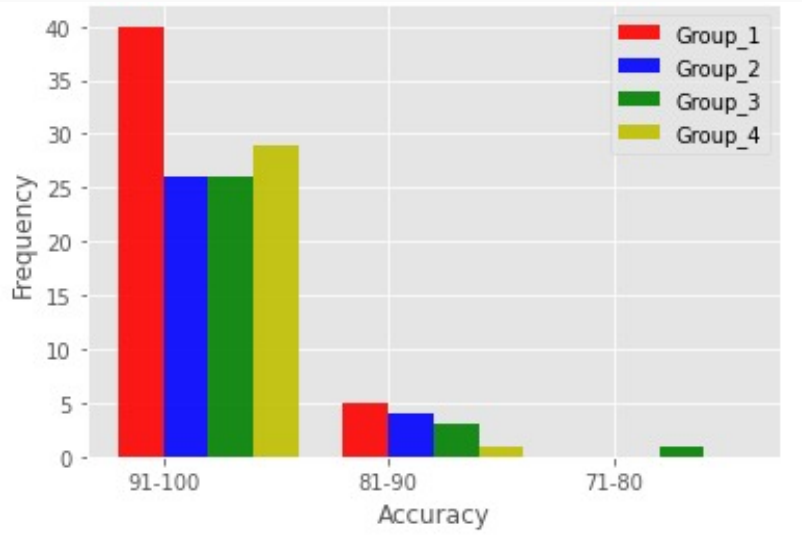
(a)

Accuracy	
Group	mean
G-1	95.22
G-2	96.167
G-3	93.47
G-4	95.93

(b)


Group_1
Group_2
Group_3
Group_4
Frequency
WPM
6-50
81-90
61-70
51-60
91-140
71-80


WPM
Group
Group_1
Group_2
Group_3
Group_4


Group_1
Group_2
Group_3
Group_4
Frequency
Accuracy
91-100
81-90
71-80

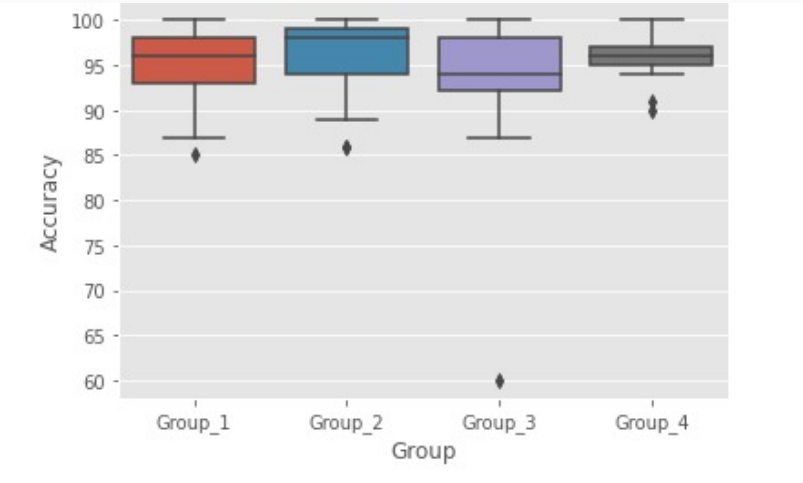

Accuracy
Group
Group_1
Group_2
Group_3
Group_4

Figure 15

Analyzing Data

CI for estimation of Population Means

Formulas

If $X_1, X_2, ..., X_n$ are normally distributed with unknown mean μ and variance σ^2, then a $(1 - \alpha)100\%$ CI for the population mean μ is:

$$\left(\bar{X} - t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right),\ \bar{X} + t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)\right)$$

CI for estimation of Population Means

95% Confidence Interval							
	UG	PHD	Masters		UG	PHD	Masters
	WPM				Accuracy		
mean	[50.40, 62.23]	[49.26, 58.97]	[48.82, 59.52]		[93.30, 96.46]	[94.88, 96.99]	[94.10, 96.46]

95% confidence interval:				
	Male	Female	Male	Female
	WPM		Accuracy	
pop_mean	[56.45,67.52]	[44.83,53.53]	[94.87,96.56]	[92.72,97.02]

WPM	
95% CI	
Group	mean
G-1	(58.63, 73.95)
G-2	(41.23, 49.43)
G-3	(48.32, 60.42)
G-4	(49.42, 57.72)

Accuracy	
95% CI	
Group	mean
G-1	(94.19, 96.26)
G-2	(94.60, 97.74)
G-3	(90.72, 96.21)
G-4	(95.08, 96.79)

CI for estimation of Difference in Population Means

Two sampled pooled t-interval

if $\left(\frac{1}{4} < \frac{S_X^2}{S_Y^2} < 4\right)$

$X_1, X_2, ..., X_n \sim N(\mu_1, \sigma^2)$ and $Y_1, Y_2, ..., Y_m \sim N(\mu_2, \sigma^2)$ are independent random samples, then a $(1-\alpha)100\%$ CI for the difference in the population means, $\mu_1 - \mu_2$ is:

$$\left((\bar{X} - \bar{Y}) - t_{\alpha/2,n+m-2} S_P \sqrt{\frac{1}{n} + \frac{1}{m}},\ (\bar{X} - \bar{Y}) + t_{\alpha/2,n+m-2} S_P \sqrt{\frac{1}{n} + \frac{1}{m}}\right)$$

where

$$S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

CI for estimation of Difference in Population Means

Welch's t-interval

if $\left(\frac{S_X^2}{S_Y^2} < \frac{1}{4} \text{ or } 4 < \frac{S_X^2}{S_Y^2}\right)$

$X_1, X_2, ..., X_n \sim N(\mu_1, \sigma_X^2)$ and $Y_1, Y_2, ..., Y_m \sim N(\mu_2, \sigma_Y^2)$ are independent random samples, then a $(1 - \alpha)100\%$ CI for the difference in the population means, $\mu_1 - \mu_2$ is:

$$\left((\bar{X} - \bar{Y}) - t_{\alpha/2,r}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}},\ (\bar{X} - \bar{Y}) + t_{\alpha/2,r}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}\right)$$

where

$$r = \text{integer part of } \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{(S_X^2/n)^2}{n-1} + \frac{(S_Y^2/m)^2}{m-1}}$$

CI for estimation of Difference in Population Means

	WPM	Accuracy
	Difference of Mean	Difference of mean
group 1,2	(12.39, 29.54)	(-0.83, 2.71)
group 1,3	(1.46, 22.39)	(-0.76, 4.27)
group 1,4	(4.11, 21.33)	(-0.72, 2.14)
group 2,3	(1.88, 16.19)	(-0.39, 5.79)
group 2,4	(2.52, 13.94)	(-1.52, 1.98)
group 3,4	(-6.38, 7.98)	(-6.38, 7.98)

(a)

difference of two means WPM :	[5.85,19.77]
difference of two means accuracy :	[-1.44,3.14]

(b)

difference of mean WPM (UG-PHD):	[-6.48, 10.87]		difference of mean Accuracy (UG-PHD):	[-0.82, 2.93]
difference of mean WPM (Masters-PHD):	[-7.06, 7.16]		difference of mean Accuracy (Masters-PHD):	[-0.90, 2.12]
difference of mean WPM (UG-Masters):	[-8.62, 8.72]		difference of mean Accuracy (UG-Masters):	[-1.58, 2.89]

(c)

CI for estimating Population Variance

Formulas

If $X_1, X_2, ..., X_n$ are normally distributed and $a = \chi^2_{1-\alpha/2,n-1}$, $b = \chi^2_{\alpha/2,n-1}$, then a $(1-\alpha)100\%$ CI for the population variance σ^2 is:

$$\left(\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right)$$

CI for estimating Population Variance

95% Confidence Interval							
	UG	PHD	Masters		UG	PHD	Masters
	WPM				Accuracy		
variance	[377.33, 781.23]	[121.37, 328.33]	[158.66, 416.28]		[26.87, 55.63]	[5.73, 15.52]	[7.75, 20.35]

95% confidence interval:				
	Male	Female	Male	Female
	WPM		Accuracy	
pop_SD	[25.57,40.33]	[10.97,17.3]	[3.85,6.07]	[5.41,8.54]

WPM	
95% CI	
Group	variance
G-1	(445.93,1038.25)
G-2	(76.52, 218.03)
G-3	(166.50, 474.41)
G-4	(78.28, 223.05)

Accuracy	
95% CI	
Group	variance
G-1	(8.16, 18.99)
G-2	(11.20, 31.92)
G-3	(34.19, 97.43)
G-4	(3.32, 9.46)

CI for estimating ratio of population variance

Formulas

If $X_1, X_2, ..., X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, ..., Y_m \sim N(\mu_Y, \sigma_Y^2)$ are independent samples and $c = F_{\alpha/2}(m-1, n-1)$,
$d = F_{1-\alpha/2}(m-1, n-1)$,

$$\left(c\frac{S_X^2}{S_Y^2},\ d\frac{S_X^2}{S_Y^2} \right)$$

CI for estimating ratio of population variance

	WPM	Accuracy
	CI for Ratio of Variances	CI for Ratio of Variances
group 1,2	(3.13,9.31)	(0.39,1.163)
group 1,3	(1.44,4.28)	(0.13,0.38)
group 1,4	(3.06,9.095)	(1.32,3.92)
group 2,3	(0.24,1.73)	(0.17,1.23)
group 2,4	(0.52,3.68)	(1.79,12.69)
group 3,4	(0.24,3.63)	(0.05,0.75)

(a)

ratio of population variance WPM	[1.66,5.15]
ratio of population variance accuracy	[0.15,0.48]

(b)

CIs for ratio of 2 variances WPM (UG-PHD):	[1.47, 5.04]		ratio of 2 variances Accuracy (UG-PHD):	[2.21, 7.60]
CIs for ratio of 2 variances WPM (Masters-PHD):	[0.38, 1.55]		ratio of 2 variances Accuracy (Masters-PHD):	[0.37, 1.50]
CIs for ratio of 2 variances WPM (UG-Masters):	[0.26, 0.87]		ratio of 2 variances Accuracy (UG-Masters):	[0.18, 0.60]

(c)

Hypothesis Testing

Definition

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

Test for $\mu_1 - \mu_2$: Independent Samples, Unequal Variances

Left-tailed test:

Null Hypothesis (H_0): $\mu_1 \geq \mu_2$

Alternative Hypothesis (H_a): $\mu_1 < \mu_2$

The **Test Statistic (TS)** is:

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Now,

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c)^2(n_1 - 1) + c^2(n_2 - 1)} \quad \text{where} \quad c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Test for $\mu_1 - \mu_2$: Independent Samples, Unequal Variances

Left-tailed test contd.

df represents degree of freedom.

Using p-value approach, we find the p-value for the test statistic (TS) by using t-distribution table and if

$$\text{p-value} \leq \alpha : \text{Reject } H_0$$
$$\text{p-value} > \alpha : \text{Fail to reject } H_0$$

Test for σ_1^2 and σ_2^2

Two-tailed test

$$\text{Null Hypothesis } (H_0): \sigma_1^2 = \sigma_2^2$$
$$\text{Alternative Hypothesis } (H_a): \sigma_1^2 \neq \sigma_2^2$$

The **Test Statistic (TS)** is:

$$F = \frac{S_1^2}{S_2^2}$$

For a significance level α, with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$, if

$$F \leq F_{1-\alpha/2,df_1,df_2} \text{ or } F \geq F_{\alpha/2,df_1,df_2}\text{: Reject } H_0$$
$$F_{1-\alpha/2,df_1,df_2} < F < F_{\alpha/2,df_1,df_2} \text{ : Fail to reject } H_0$$

Gender-Based Hypothesis Testing

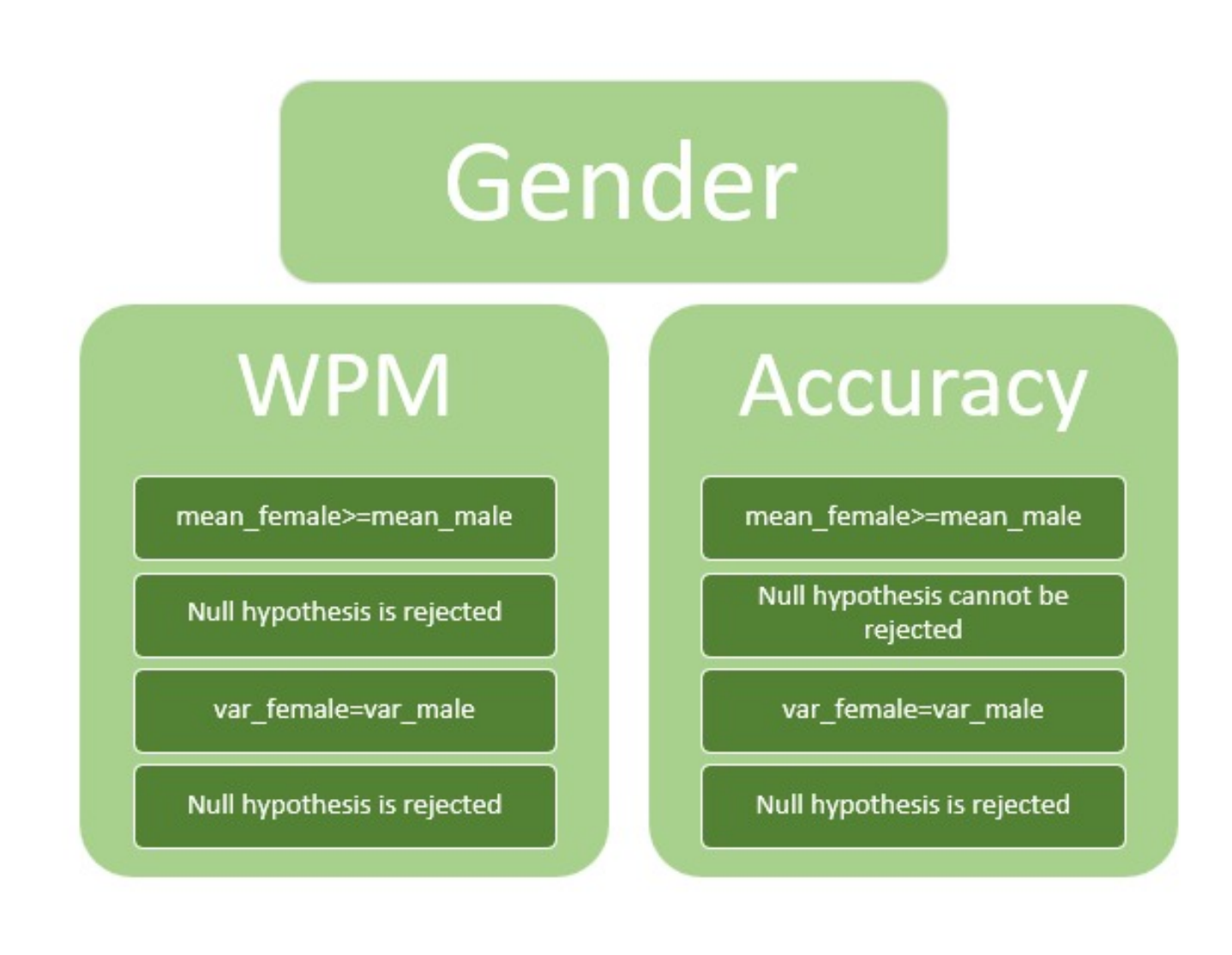


Figure 18

Batch-Based Hypothesis Testing

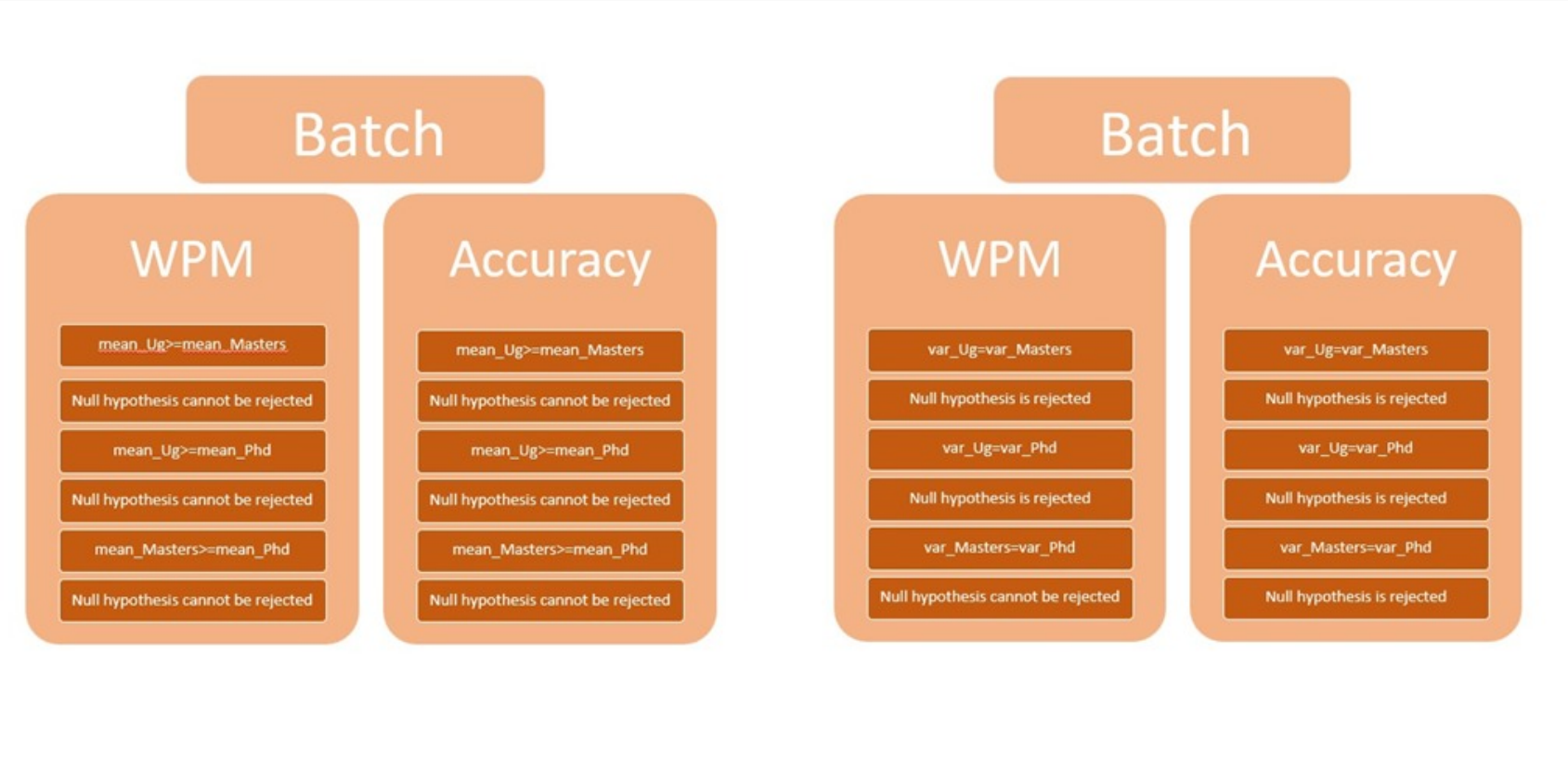


Figure 19

Branch-Based Hypothesis Testing

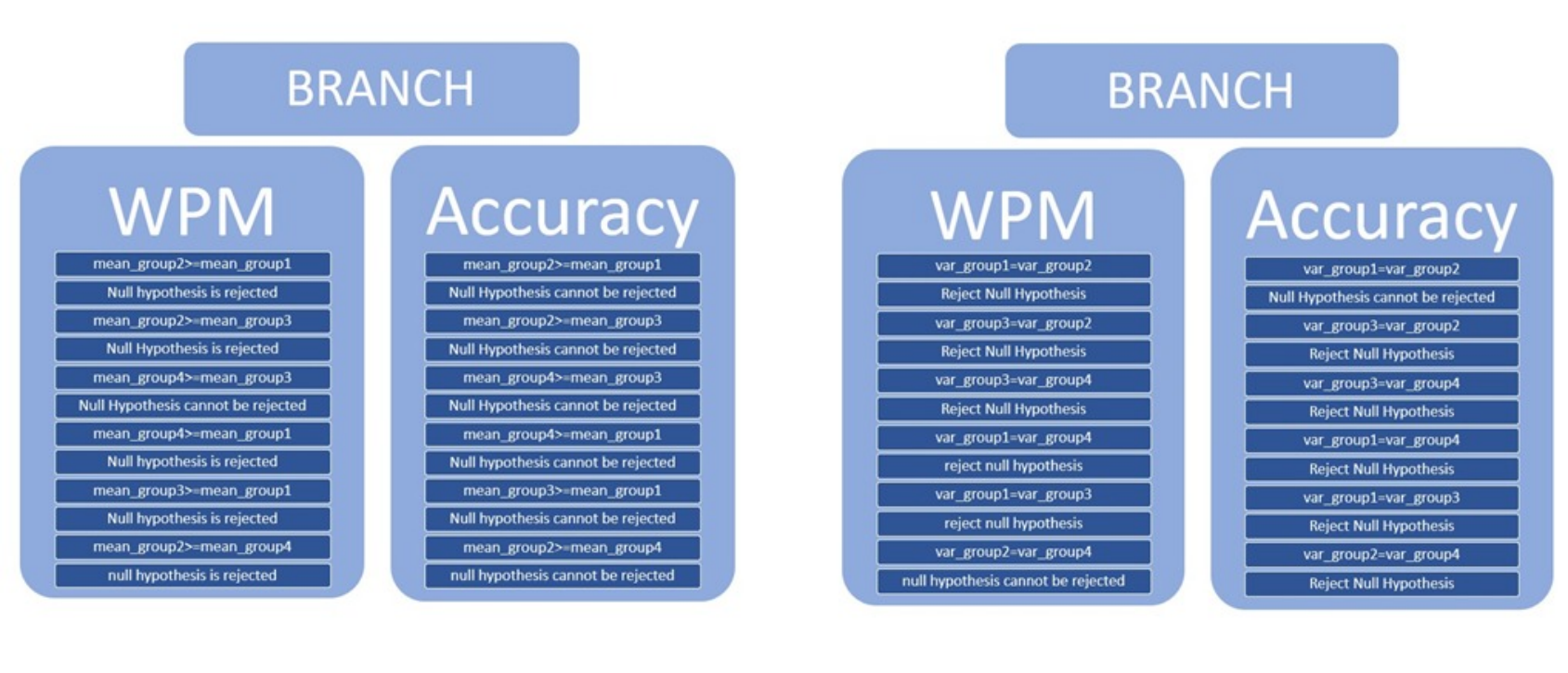


Figure 20

Conclusion

Conclusion

We draw the following conclusions:

- Since we performed an observational study, we can only ascertain correlation between the dependent and independent variables.
- In terms of words per minute, males have been found to perform better than their female counterparts. The spread[1] of data is not similar in both cases.

[1] By spread, it means the comparison of variances of the groups

Conclusion

- In terms of wpm, Group 1(CSE, MA, AI, ES) performed the best .Group 2 (CH, CE, MSME, BME,BTE) had the least typing speed. No concrete conclusion can be drawn about the order between Group 3(Physics, ME, EE, EP) and Group 4(LA, Design, Others). Spread of all groups is different but group 1 and group 2 can not be conclusively compared.
- No clear correlation was found between the typing speed of an individual and the course they are enrolled in.UG's have different spread than Masters and PhD's. Masters and PhD's can not be conclusively compared.
- No such conclusion can be drawn about accuracy in any of the above stated cases.

THANK YOU!

▸ Google Sheet Link with all the data related to the study