Applied Analytics - Assignment 2

Analysis of Employee Neuroticism and Turnover Tendency

Huu Thien Ngoc Phung - s3643808

Last updated: 15 October, 2021

Introduction

- In conducting business, high employee turnover rate can become costly for organisations because it takes time to recruit and train new people to fill the vacancies.
- For years, the Big Five personality traits including openness, conscientiousness, extraversion, agreeableness, and neuroticism have been widely used in human resource management for predicting employee performance (Barrick & Parks & Mount, 2005).
- This report will study the difference in neuroticism level of individuals leaving and individuals staying with their employers to confirm if neuroticism can be used as an indicator in predicting employee turnover.

Problem Statement

- The purpose of this analysis is to investigate if there is a statistical difference in the anxiety level (measurement for neuroticism) of employees who left their organisations and employees who stayed with their organisations.
- The investigation uses R programming software to analyse an open data set from Kaggle about employee turnover. Necessary data cleaning techniques are performed to prepare the data for the statistical analysis.
- As a first step of the analysis, the bivariate visualisation technique is applied to learn about the distributions of anxiety levels of employees in each turnover event groups (stayed and left).
- Next, summary statistics of neuroticism levels is produced for an in-depth comparison between the two groups.
- After that, the analysis applies two-tailed two-sample t-test to confirm if there is a statistical difference between the two groups' anxiety levels.
- In preparation for the above hypothesis test, the report also checks for the normality assumption and homogeneity of variance for both groups before conducting the test.
- At the end of the analysis is a discussion of the findings and recommendation for future works.

Data

- This analysis uses the data set 'turnover.csv' collected by Edward Babushkin, a Russian people analyst (Babushkin, 2020).
- The data set containing 16 variables with the records of 1,129 employees working in different industries including their score for each of the big five personality traits.
- For this investigation, we subset the data set with only 3 variables of interest. Their column names and data types are:
  - ***event*** *[factor]: the employee turnover event, '0' means the employee stayed with the company, '1' means the employee quitted.*
  - ***industry*** *[character]: the industry that the employee was working in.*
  - ***anxiety*** *[numeric]: anxiety score, used to measure personality trait of neuroticism. Values are scaled from 1 to 10, the higher the score, the less emotional stability the employee appeared to be.*

```
turnoverfull <- read_csv("turnover.csv")
turnover <- turnoverfull %>% select(event, industry, anxiety)
head(turnover)
```

event <dbl>	industry <chr>	anxiety <dbl>
1	Banks	7.1
1	Banks	7.1
1	PowerGeneration	4.8
1	PowerGeneration	2.5
1	Retail	7.1
1	manufacture	5.6

6 rows

Data Cont.

- Convert the data type of 'event' from numeric to 'factor' with 2 levels: 'Stayed' for employees who stayed with their employers, and 'Left' for employees who left their jobs.

```
turnover$event %>% unique()
```

```
## [1] 1 0
```

```
turnover$event <- turnover$event %>%
  factor(levels = c(0,1),labels = c("Stayed", "Left"))
levels(turnover$event)
```

```
## [1] "Stayed" "Left"
```

- Checking for missing values and special values: no such values found.

```
colSums(is.na(turnover))
```

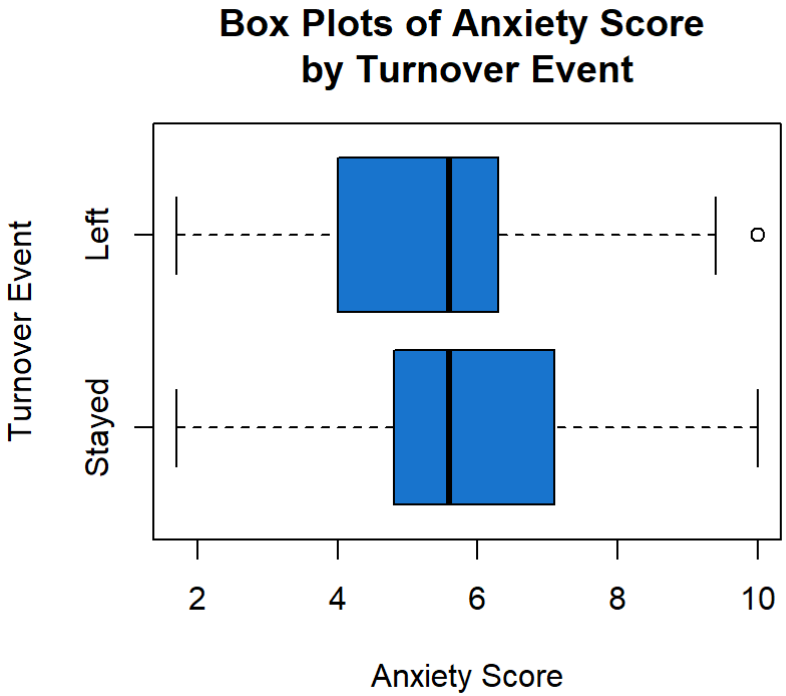
```
##    event industry  anxiety
##        0        0        0
```

```
is.special <- function(x){
if (is.numeric(x)) (is.infinite(x) | is.nan(x))}
sum(is.special(turnover$anxiety))
```

```
## [1] 0
```

Descriptive Statistics and Visualisation

```
bp1 <- boxplot(turnover$anxiety ~ turnover$event, col="dodgerblue3",
               horizontal = TRUE, xlab = "Anxiety Score", ylab = "Turnover Event",
               main = "Box Plots of Anxiety Score \nby Turnover Event")
```



As we are investigating the anxiety score for each of the event groups (Stayed vs Left), the bivariate boxplots technique is applied to compare the employee anxiety score distribution for each event group and detect potential outliers using Tukey's method of outlier detection.

Interpretation of the Box Plots

- The two box plots shows that even though the first quartile and third quartile of Left group anxiety score is lower than Stay group's, the median anxiety score in both groups are similar.
- The interquartile ranges of the anxiety score in both groups as shown by the lengths of the boxes are reasonably similar.
- However, the whisker of Stay group is larger than that of Left group, indicated a wider distribution in anxiety score in Stay group.
- Anxiety score of Left group appear to be left-skew, and that of Stay group appear to be right-skew.
- There are 3 detected outliers lying above the Left group upper fence, all with value of 10.

```
bp1$out
```

```
## [1] 10 10 10
```

Descriptive Statistics Cont.

Handling outliers

- First, subset the 'turnover' data frames into 2 new data frames which are 'Stayed' containing all employees who stayed with their organization and 'Left' containing all employees who quit from their job.

```
Stayed <- turnover %>% filter(event=="Stayed")
Left <- turnover %>% filter(event=="Left")
dim(Left)
```

```
## [1] 571   3
```

- As there are only 3 outliers (as shown from the box plot) out of 571 observations in Left group, removing these outliers would be a safe approach without biasing the analysis.
- Next, excluding 3 observations considered as Tukey's method detected outliers from 'Left' data frame.

```
Left <- Left[!Left$anxiety > quantile(Left$anxiety, 0.75) +
                       1.5*IQR(Left$anxiety), ]
dim(Left)
```

```
## [1] 568   3
```

Descriptive Statistics Cont.

After removing outliers

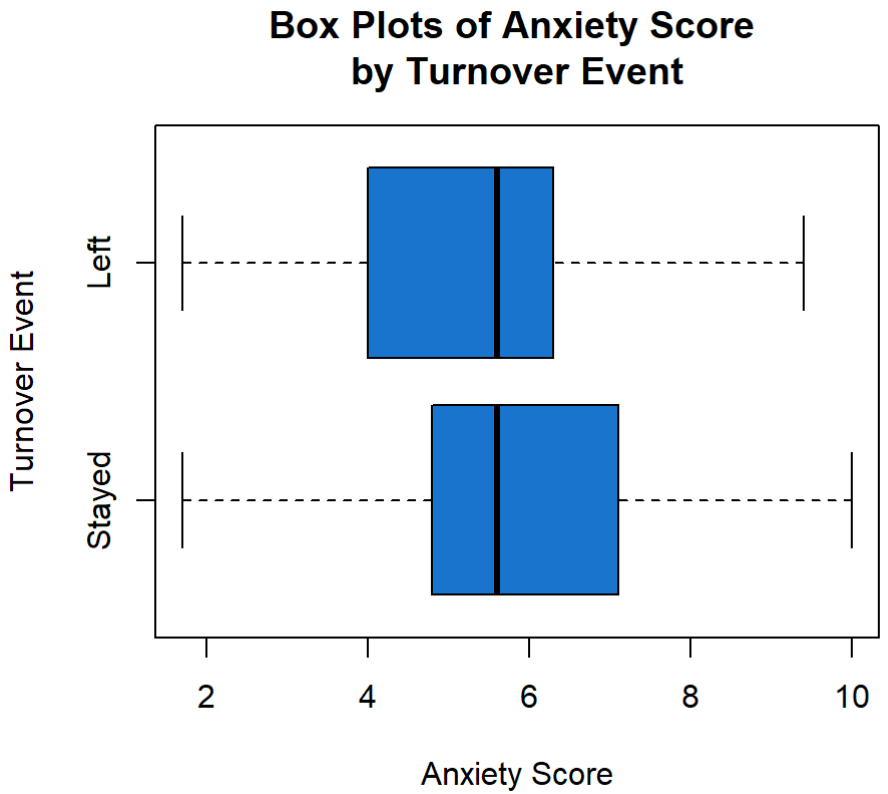
- Combine the 'Stayed' and 'Left' into new data frame 'turnover_clean'.

```
turnover_clean <- bind_rows(Stayed, Left)
str(turnover_clean)
```

```
## tibble [1,126 x 3] (S3: tbl_df/tbl/data.frame)
##  $ event   : Factor w/ 2 levels "Stayed","Left": 1 1 1 1 1 1 1 1 1 1 ...
##  $ industry: chr [1:1126] "Retail" "Retail" "Retail" "manufacture" ...
##  $ anxiety : num [1:1126] 5.6 7.1 7.1 5.6 7.1 7.1 7.1 5.6 6.3 7.9 ...
```

- Draw new Box Plots of Anxiety Score by Turnover Event from the cleaned data, which is no longer showing outliers.

```
bp2 <- boxplot(turnover_clean$anxiety ~ turnover_clean$event, col="dodgerblue3",
               horizontal = TRUE, xlab = "Anxiety Score", ylab = "Turnover Event",
               main = "Box Plots of Anxiety Score \nby Turnover Event")
```



Descriptive Statistics Cont.

Summary Statistics

```
turnover_clean %>% group_by(`Turnover Event` = event) %>%
  summarise(Min = min(anxiety,na.rm = TRUE),
            Q1 = quantile(anxiety,probs = .25,na.rm = TRUE),
            Median = median(anxiety, na.rm = TRUE),
            Q3 = quantile(anxiety,probs = .75,na.rm = TRUE),
            Max = max(anxiety,na.rm = TRUE),
            Mean = round(mean(anxiety, na.rm = TRUE),2),
            IQR = IQR(anxiety,na.rm = TRUE),
            SD = round(sd(anxiety, na.rm = TRUE),2),
            n = n(),
            Missing = sum(is.na(anxiety))) -> tab1
tab1 %>% kbl(caption = "Table 1: Summary Statistics of Anxiety Score") %>%
  kable_classic_2(full_width = F) %>%
  row_spec(0,bold=TRUE, color="brown") %>%
  column_spec(1, bold=TRUE, color="brown")
```

Table 1: Summary Statistics of Anxiety Score

Turnover Event	Min	Q1	Median	Q3	Max	Mean	IQR	SD	n	Missing
Stayed	1.7	4.8	5.6	7.1	10.0	5.77	2.3	1.68	558	0
Left	1.7	4.0	5.6	6.3	9.4	5.54	2.3	1.70	568	0

- Table 1 showed summary statistics of anxiety score for each turnover event group.
- The minimum value and median value of both groups are similar at 1.7 and 5.6 respectively.
- Stayed group has higher maximum value of anxiety score than Left group (10>9.4).
- The interquartile range of both group are the same which is 2.3. However, the first quartile Q1 and third quartile Q3 of Stayed group are 4.8 and 7.1 respectively are both higher than the corresponding statistics in Left group, which are 4 and 6.3 respectively.
- The mean anxiety score of Stayed group is 5.77, higher than the mean of Left group which is 5.54.
- The standard deviation of Stayed group is 1.68, very close to the standard deviation of Left group which is 1.7.
- For Stayed group, the mean score (5.77) is higher than the median score (5.6), suggested that the distribution of anxiety score in Stayed group is right skewed.
- For Left group, the mean score (5.54) is lower than the median score (5.6), suggested that the distribution of anxiety score in Left group is left skewed.

Hypothesis Testing

- From the summary statistics and the box plots, it seems that employees who stayed with their organisations tend to have a higher anxiety score than those who left their organizations.
- **Two-tailed two-sample t-test** is applied in order to compare the mean anxiety score of two employee groups (Stayed and Left) and check if this difference is statistically significant.
- Therefore, statistical hypotheses are:
  - ***Null Hypothesis*** H_0*: the difference between the two underlying population means of the anxiety score of employees who stayed and employees who Left is 0.*
  - ***Alternative Hypothesis*** H_A*: the difference between the two underlying population means of the anxiety score of employees who stayed and employees who quitted is not equal to 0.*

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

where μ_1 and μ_2 refer to population means of Stayed group and Left group respectively.

- Before conducting the two-sample t-test, the **normality assumption** and **variance homogeneity assumption** need to be checked.

Testing the Assumption of Normality for Stayed Group

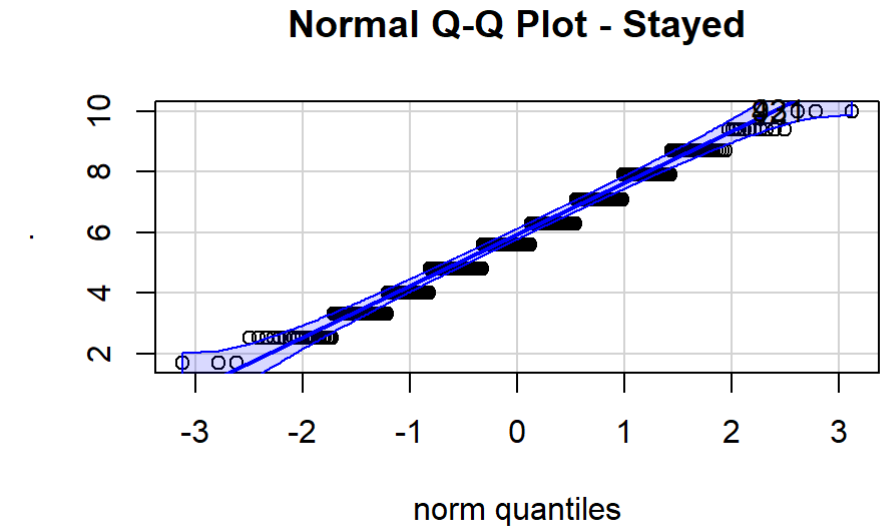
- Normal Q-Q Plot & Histogram of Anxiety Score for Stayed Group.

```
par(mfrow=c(2,1))
Stayed$anxiety %>% qqPlot(dist="norm", main = "Normal Q-Q Plot - Stayed")
```

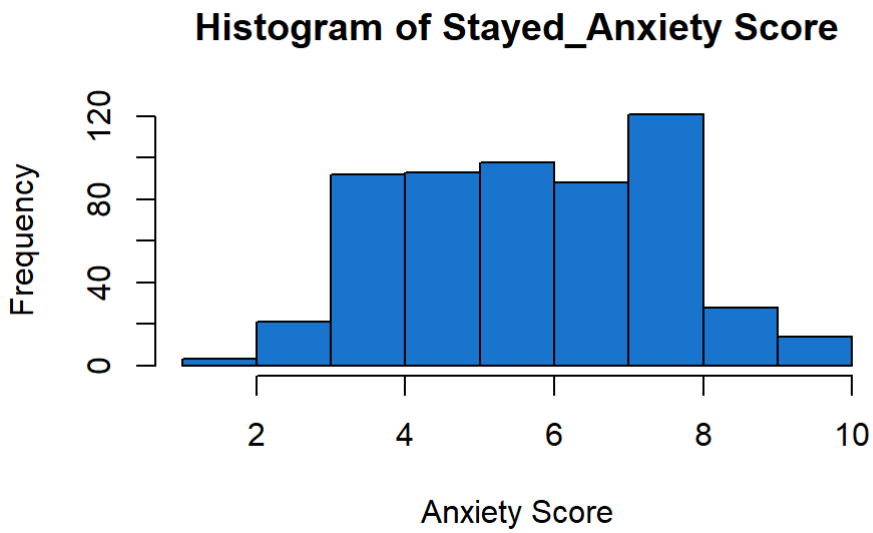
```
## [1]  93 421
```

```
Stayed$anxiety %>% hist(col="dodgerblue3",  xlab="Anxiety Score",
                        main = "Histogram of Stayed_Anxiety Score")
```

```
par(mfrow=c(1,1))
```



- The Q-Q plot suggested that the anxiety score distribution for Stayed group is not normal as there are many points falling outside, mostly to the right, of the 95% CI for the normal quantiles.
- The histogram also suggested that this is a right skew distribution, which matches with the findings from the summary statistics.
- According to the the Central Limit Theorem (CLT), when the sample size is large ($n > 30$), the sampling distribution of a mean is approximately normal regardless of the variable's underlying population distribution (Baglin, 2021).
- From the summary statistics, the sample size of Stayed group data is $n = 558$, therefore, thanks to the CLT, we can ignore the non-normality issue for the Stayed group data and proceed with the two-sample t-test.



Testing the Assumption of Normality for Left Group

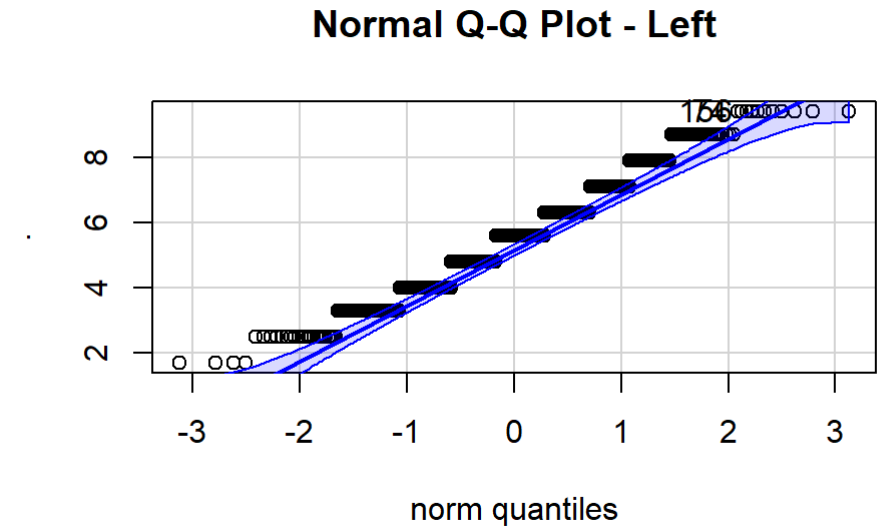
- Normal Q-Q Plot & Histogram of Anxiety Score for Left Group.

```
par(mfrow=c(2,1))
Left$anxiety %>% qqPlot(dist="norm", main = "Normal Q-Q Plot - Left")
```

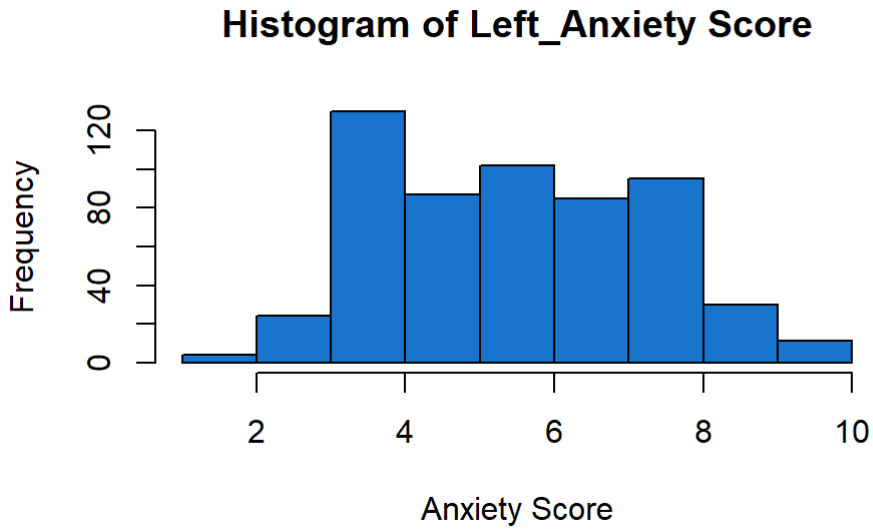
```
## [1]  74 156
```

```
Left$anxiety %>% hist(col="dodgerblue3", xlab="Anxiety Score",
                      main = "Histogram of Left_Anxiety Score")
```

```
par(mfrow=c(1,1))
```



- The Q-Q plot suggested that the anxiety score distributions for Left group is not normal as there are many points falling outside, mostly to the left, of the 95% CI for the normal quantiles.
- The histogram also suggested that this is a left skew distribution, which matches with the findings from the summary statistics.
- Even though the normality assumption is violated, as the sample size is large, $n = 568$, thanks to the CLT, we can ignore the non-normality issue for the Left group data and proceed with the two-sample t-test.



Hypothesis Testing Cont.

Homogeneity of Variance

Statistical hypotheses of Levene's test of equal variance for anxiety score between employees who stayed and who quitted:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_A : \sigma_1^2 \neq \sigma_2^2$$

where σ_1^2 and σ_2^2 refer to the population variance of the anxiety score of Stayed group and Left group respectively.

```
leveneTest(anxiety ~ event, data = turnover_clean)
```

	Df <int>	F value <dbl>	Pr(>F) <dbl>
group	1	0.08737192	0.7675998
	1124	*NA*	*NA*

2 rows

- The p-value for the Levene's test is $p = 0.77$.
- As $p > 0.05$, we fail to reject H_0.
- Hence, it is safe to assume equal variance.

Hypothesis Testing Cont.

Two-sample t-test - Assuming Equal Variance

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

```
t.test(anxiety ~ event, data = turnover_clean, var.equal = TRUE, alternative = "two.sided")
```

```
##
##  Two Sample t-test
##
## data:  anxiety by event
## t = 2.3733, df = 1124, p-value = 0.0178
## alternative hypothesis: true difference in means between group Stayed and group Left is not equal to 0
## 95 percent confidence interval:
##  0.04150244 0.43754370
## sample estimates:
## mean in group Stayed   mean in group Left
##             5.774910             5.535387
```

- The anxiety score difference between Stayed group and Left group estimated by the sample was $5.775 - 5.535 = 0.24$
- The two-tailed p-value is reported to be $p = 0.018$.
- The 95% CI of the difference between the means (0.24) is reported as 95% CI $[0.04\ 0.44]$.
- Using the p-value method, as $p = 0.018 < \alpha = 0.05$, we can reject H_0.
- Therefore, there is a statistically significant difference between the means of anxiety score in both groups.

Discussion

- The summary statistics and box plots suggested that even though both groups shared the same median anxiety score, the mean anxiety score of Stayed group is greater than Left group.
- A two-tailed two-sample t-test was used to test for a significant difference between the mean anxiety score of stayed employees and Left employees. While the anxiety scores for both groups demonstrated evidence of non-normality upon checking the normal Q-Q plots, the central limit theorem ensured that the t-test could be applied due to the large sample size in each turnover event group. The Levene's test of homogeneity of variance showed that it is safe to assume equal variance.
- The results of the two-sample t-test assuming equal variance found statistically significant difference between the anxiety scores of stayed employees and Left employees, $t(df = 1124) = 2.37$, $p = .018$, 95% CI for the difference in means $[0.04\ 0.44]$.
- The results of the investigation suggested that stayed employees have significant higher average anxiety score than Left employees.
- This indicated that neuroticism is likely to affect employees' likelihood to leave their jobs. The higher an individual anxiety level, the less likely they want to change their jobs.
- The scope of this report is limited to study one personality trait which is neuroticism in predicting employee turnover. It is suggested that future research can use logistic regression analysis for all five personality traits to study the likelihood of an individual quitting.

References

Babushkin, E 2020, *Employee Turnover*, data file, Kaggle, viewed 15 October 2021, https://www.kaggle.com/davinwijaya/employee-turnover.

Baglin, J 2021, 'Module 5 Sampling: Randomly Representative', lecture notes, MATH1324, RMIT University, Melbourne, viewed 15 October 2021, https://astral-theory-157510.appspot.com/secured/MATH1324_Module_05.html#Central_Limit_Theorem.

Barrick, MR, Parks, L & Mount, MK 2005, 'Self-Monitoring as a Moderator of the Relationship between Personality Traits and Performance', *Personnel Psychology*, vol. 58, no. 3, pp. 745-767.