Assignment 4.1 Student Assignment Performance Data

Larry Heckel

April 6, 2019

Assignment

Imagine you are a graduate assistant and your supervisor has asked you to look at student assignment performance data. The data file you will use is named Exams.dat.

The first three variables in the data are a percentage based on 100 points. The fourth variable is based on a score achieved out of 15 points. The stats variable are the letter grades for the Stats exam where A=1, B=2, C=3, D=4, F=5. The uni variable indicates which professor taught that section of the course. 0=Professor George, 1=Professor Jeff, and 3=Professor Tushmann.

Your supervisor has requested an appropriate statistical analysis be performed visually and numerically.

1. Perform an analysis for the entire dataset and for each group. In addition, produce an appropriate graph for each variable. Perform data transformations as appropriate and explain if the transformation you performed was beneficial or created more problems. Use R Markdown to report, critique and discuss the skewness and any significant scores found.

2. Perform a Shapiro-Wilk test for the entire dataset and for each group. Use R Markdown to report the results found along with an appropriate plot. Include an appropriate narrative throughout.

```r
# libraries to be used
library(ggplot2)
library(readr)
library(pastecs)
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
# read the file
examDataX <- read_delim("exams.dat", delim = '\t')
```

```
## Warning: Missing column names filled in: 'X7' [7]

## Parsed with column specification:
## cols(
##   exam = col_double(),
##   computer = col_double(),
##   lectures = col_double(),
##   numeracy = col_double(),
##   stats = col_double(),
##   uni = col_double(),
##   X7 = col_logical()
## )
```

```r
#set our display options for the analysis results
options(scipen=100)
options(digits=2)
```

```r
#removing the extraneous column in the data
examData <- within(examDataX, rm(X7))
View(examData)
str(examData)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 150 obs. of  6 variables:
##  $ exam    : num  18 30 40 30 40 15 36 40 63 31 ...
##  $ computer: num  54 47 58 37 53 48 49 49 45 62 ...
##  $ lectures: num  75 8.5 69.5 67 44.5 76.5 70 18.5 43.5 100 ...
##  $ numeracy: num  7 1 6 6 2 8 3 7 4 6 ...
##  $ stats   : num  1 1 1 2 2 2 2 3 3 3 ...
##  $ uni     : num  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   exam = col_double(),
##   ..   computer = col_double(),
##   ..   lectures = col_double(),
##   ..   numeracy = col_double(),
##   ..   stats = col_double(),
##   ..   uni = col_double(),
##   ..   X7 = col_logical()
##   .. )
```

First Impressions

We can see that there are 6 variables within the data set, of which 5 are evaluations (scores, grades) of some form, and the 6th is an identifier of the professor who taught the course.

Of the 5 evaluations, the first three (exam, computer, lectures) are represented as a continuous numerical percentage (out of 100), numeracy is an discrete integer raw score out of 15, and stats is a discrete integer value representing a standard letter grade, with 1=A and 5=F.

The approach will be to analyze each of the evaluations, to determine how their values are distributed and detect patterns within each. I'll also perform a Shapiro-Wilk test for normality on each of the evaluations. I will then do the same for the evaluations of each professor who taught the course, to see what effect the professor's teaching had on the evaluation results.

First let's look at the scores for the exam evaluation

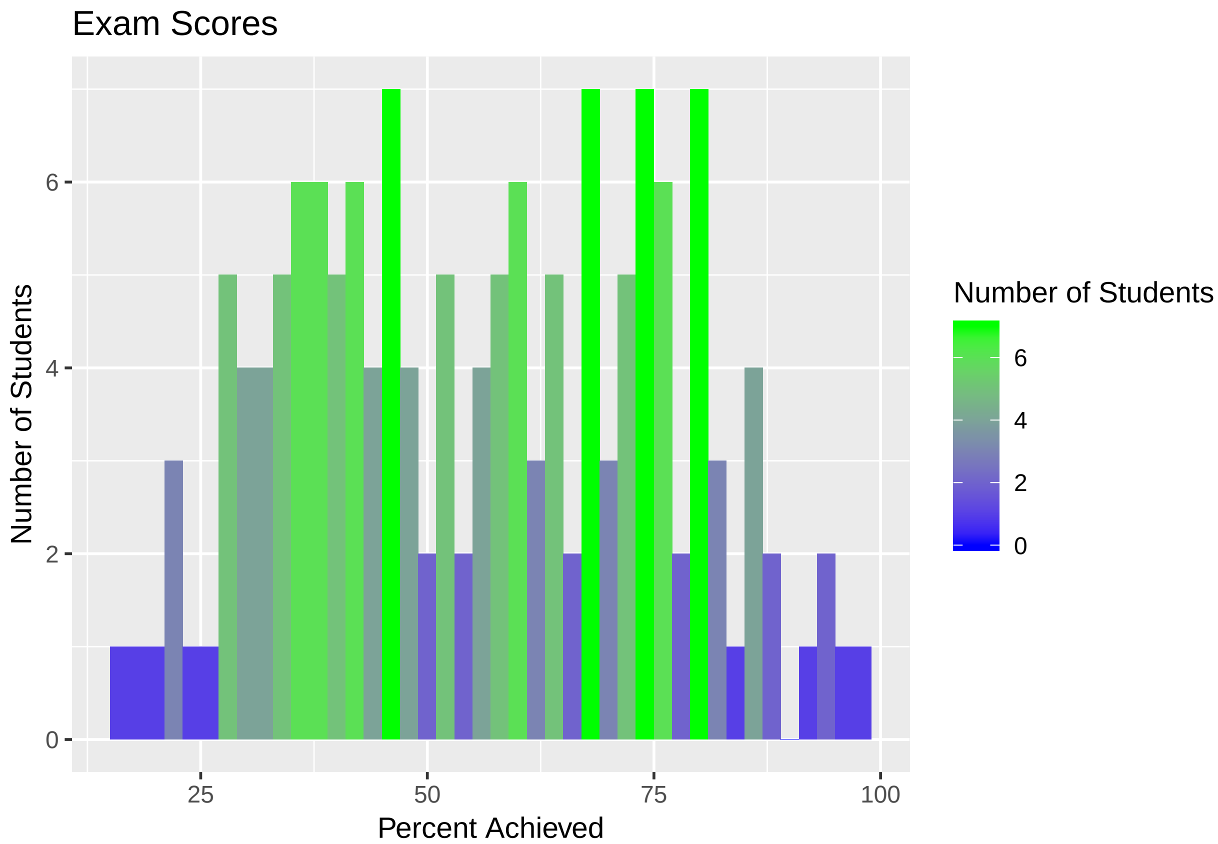
```r
# Analyze the entire data set.
examStats     <- stat.desc(examData$exam, norm = TRUE)
examStats
```

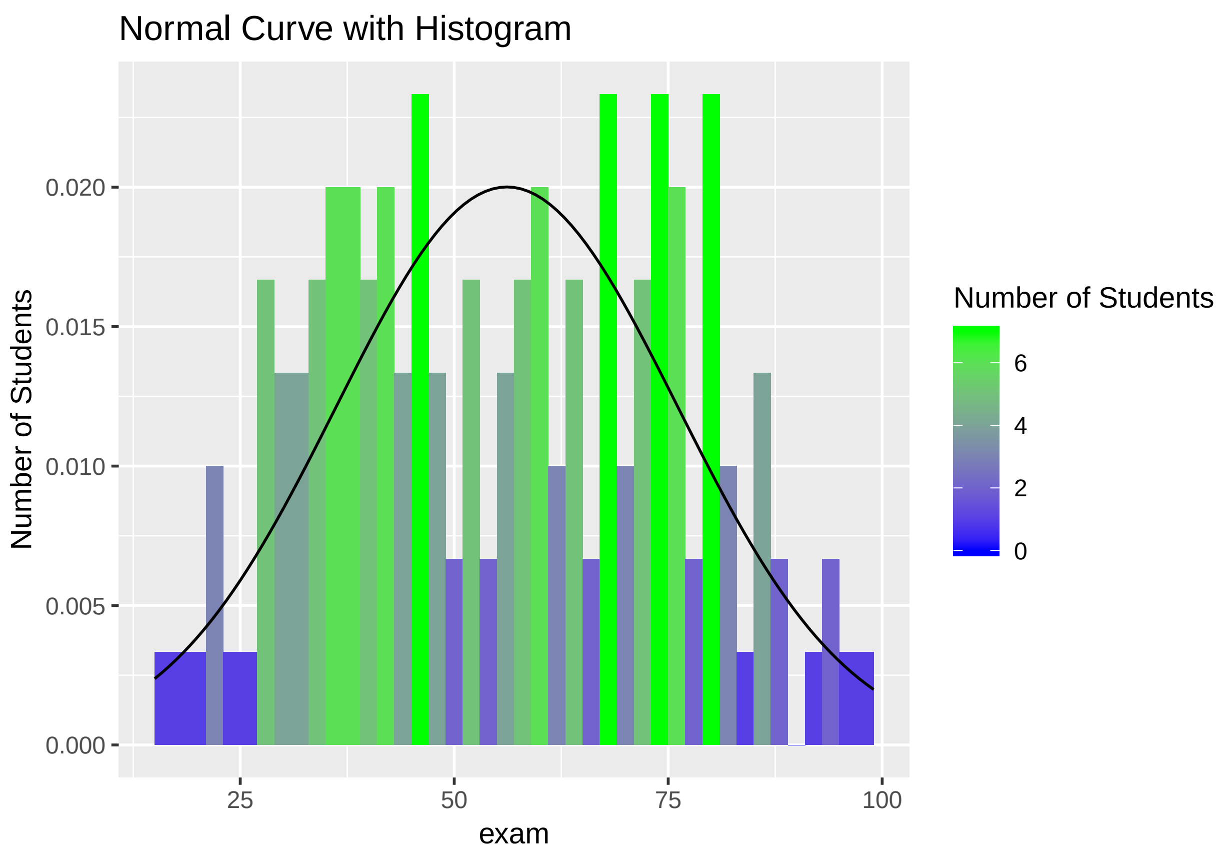
```
##      nbr.val     nbr.null       nbr.na          min          max
##     150.0000       0.0000       0.0000      15.0000      99.0000
##        range          sum       median         mean      SE.mean
##      84.0000    8424.0000      56.0000      56.1600       1.6280
## CI.mean.0.95          var      std.dev     coef.var     skewness
##       3.2170     397.5581      19.9389       0.3550       0.0614
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
```

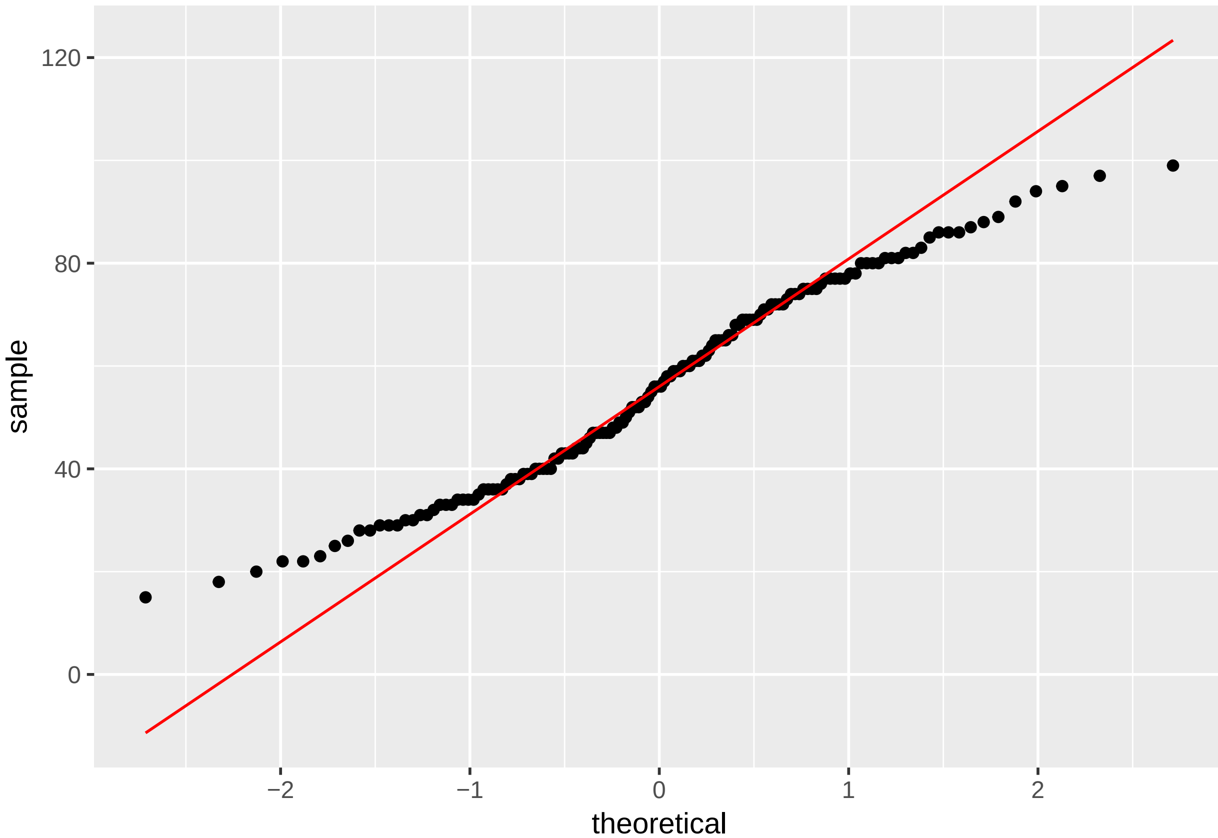
```
##       0.1550      -1.0246      -1.3016       0.9741       0.0062
```

```r
shapiro.test(examData$exam)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examData$exam
## W = 1, p-value = 0.006
```




Normal Curve with Histogram
Number of Students
0.000
0.005
0.010
0.015
0.020
25
50
75
100
exam
Number of Students
0
2
4
6



Impressions of the exam results

Without examining the numerical statistics, we can see from the first histogram that the exam results are not normally distributed. The basic histogram at first glance appears to show 3 data peaks, one each centering around 35%, 60%, and 75%, leading to a supposition that there might be use in dividing the data up by the professor, to see if that is what might be driving the data peaks.

In looking at the numerical statistics, nothing really stands out to change my initial visual impressions. The mean and median are essentially the same, meaning that the data is evenly distributed, and the skewness measures confirm that. The standard deviation of almost 20 confirms that the results are widely distributed, with a range of 84 (out of 100 possible).

For the Shapiro-Wilk test, we evaluate the p-value to determine normality. If the p-value < .05, then we conclude that the distribution of the data deviates from a normal distribution. The Shapiro-Wilk test p-value of 0.006182 confirms that the exam distribution is not normal.

Next let's look at the scores for the computer evaluation

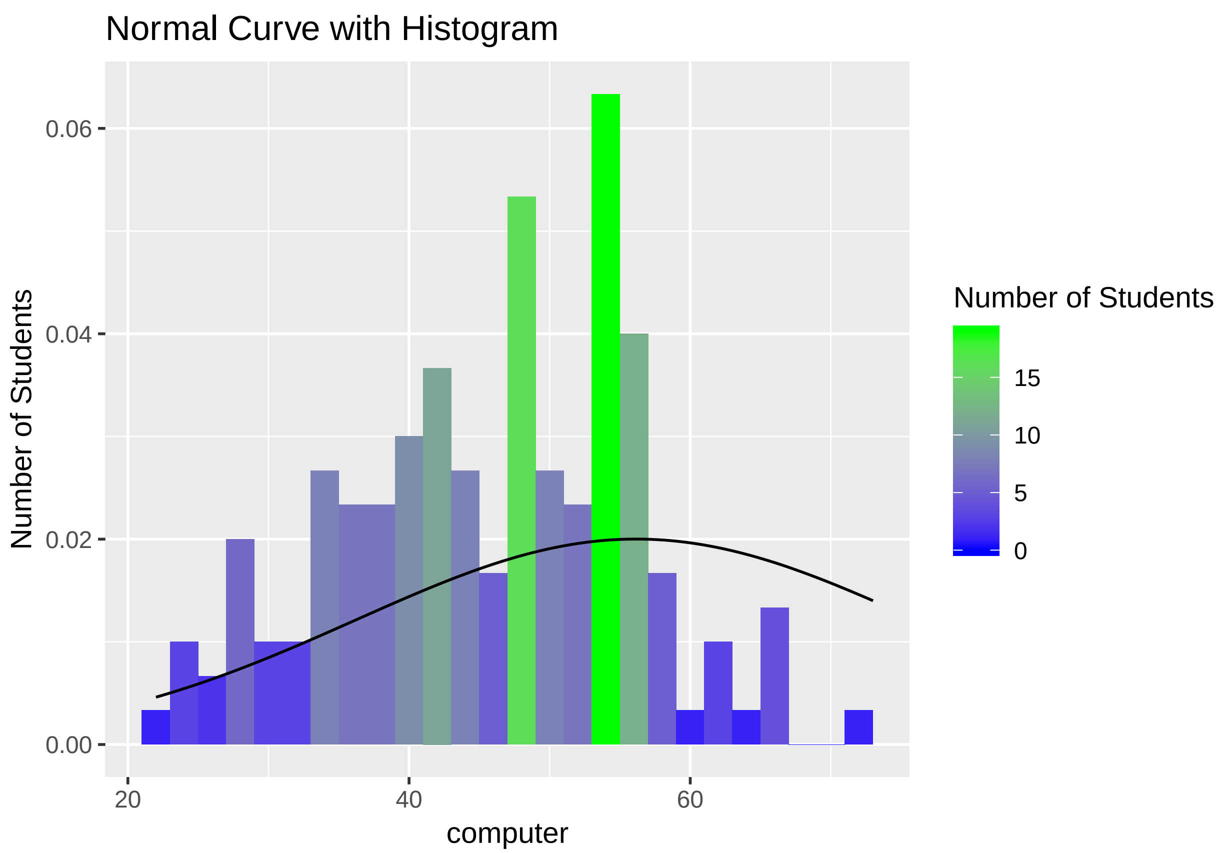
```
###################
computerStats <- stat.desc(examData$computer, norm = TRUE)
computerStats
```

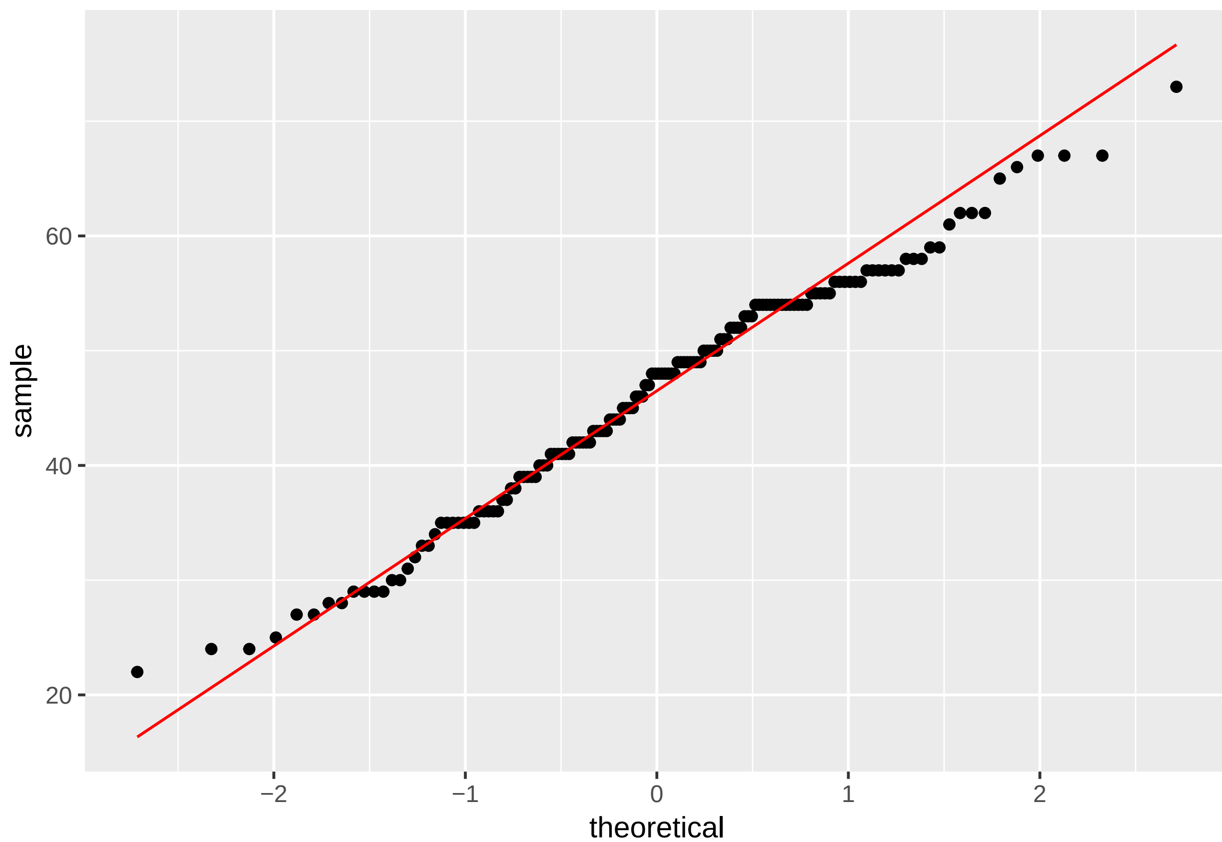
```
##      nbr.val     nbr.null       nbr.na          min          max
##       150.00         0.00         0.00        22.00        73.00
##        range          sum       median         mean      SE.mean
##        51.00      6934.00        48.00        46.23         0.85
## CI.mean.0.95          var      std.dev     coef.var     skewness
```

```
##          1.67        107.21         10.35          0.22         -0.16
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##         -0.39         -0.48         -0.62          0.99          0.11
```

```r
shapiro.test(examData$computer)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examData$computer
## W = 1, p-value = 0.1
```


Normal Curve with Histogram
Number of Students
0.00
0.02
0.04
0.06
20
40
60
computer
Number of Students
0
5
10
15



Impressions of the computer results

The histogram for the computer evaluation show a much tighter data distribution, with scores in general lower than the exam results. The students clearly did not do as well, overall, on this evaluation, as they did on the exam evaluation.

The results again show 3 fairly distinct peaks in the distribution, one each around 42%, 48%, and 54%. Similar to the conclusion for the exam results, it leads to the supposition that there might be a way of subsetting the data to yield better insights.

The lower end of the range, from around 25%-45%, has a high number of results, reflecting that overall, the results on this evaluation were not very good.

The numerical statistics again confirm what we are seeing visually. The Mean and Median are in the high 40's, meaning that the students overall did poorly on the evaluation.

The Shapiro-Wilk test p-value of 0.1103 confirms that the exam distribution is normal.

Next let's look at the scores for the lectures evaluation

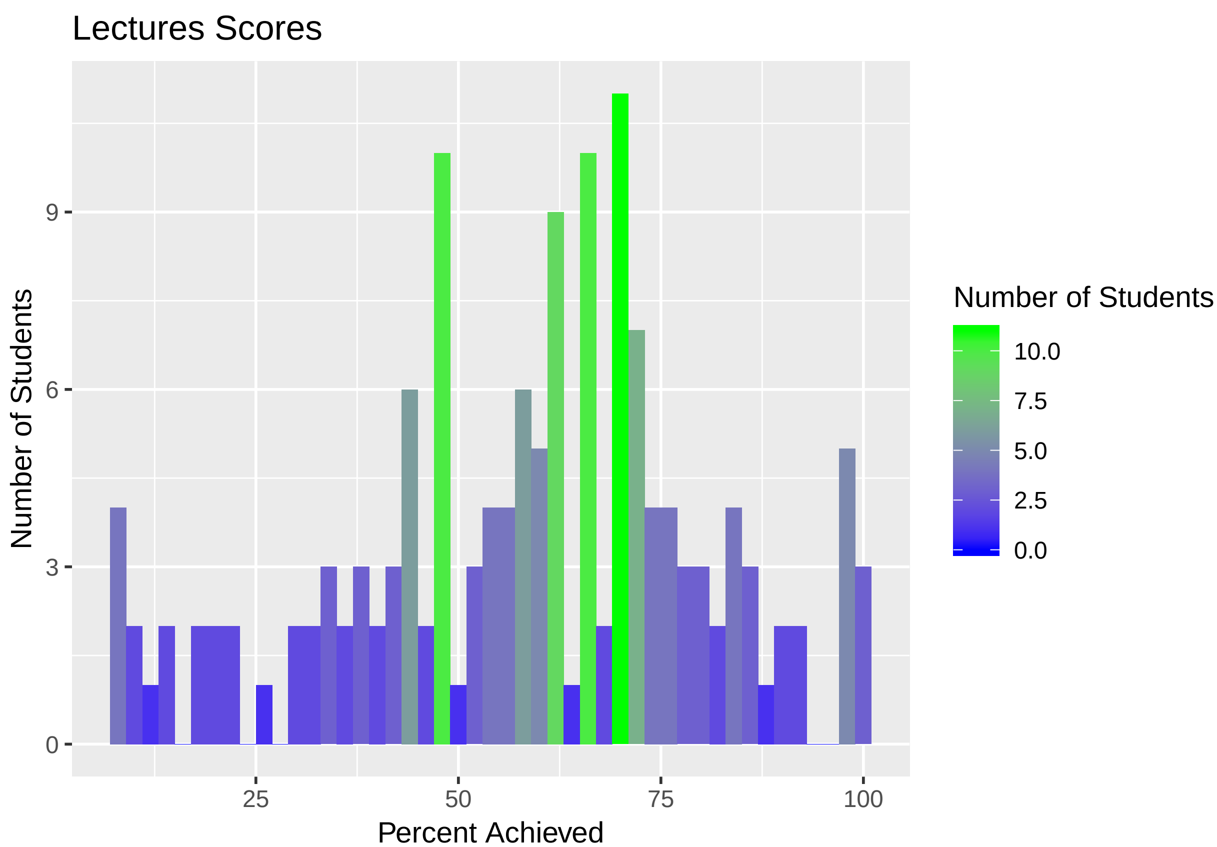
```
#########################
lecturesStats <- stat.desc(examData$lectures, norm = TRUE)
lecturesStats
```

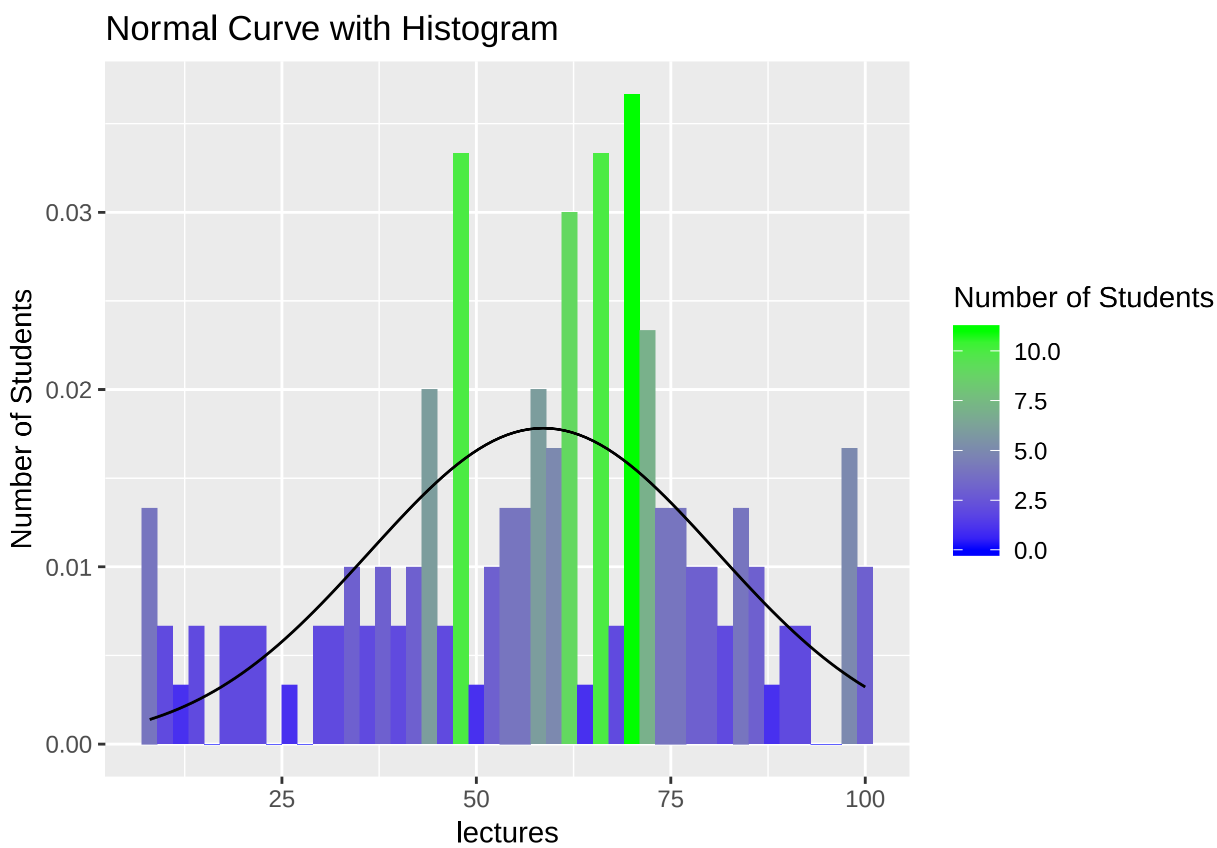
```
##      nbr.val     nbr.null       nbr.na          min          max
##     150.0000       0.0000       0.0000       8.0000     100.0000
##        range          sum       median         mean      SE.mean
##      92.0000    8789.5000      61.5000      58.5967       1.8276
```

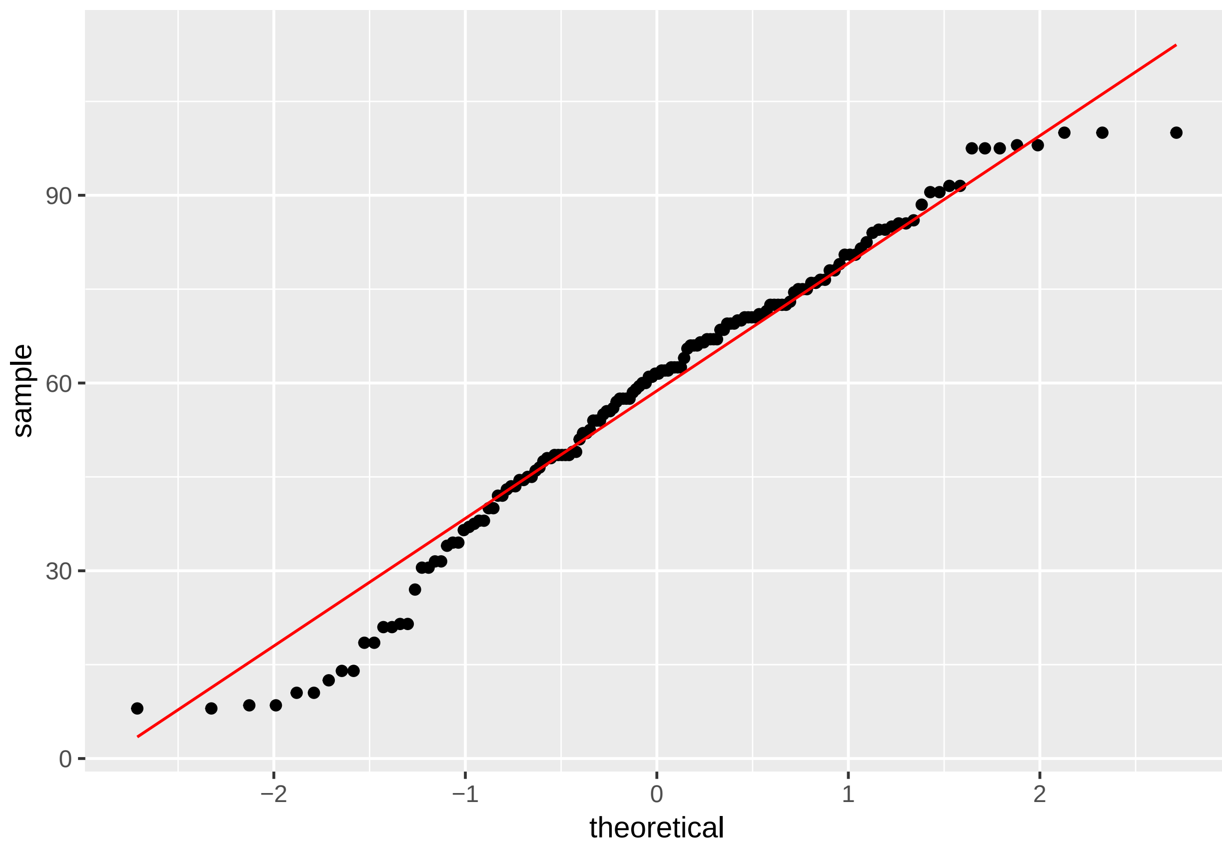
```
## CI.mean.0.95          var      std.dev     coef.var     skewness
##       3.6114     501.0325      22.3838       0.3820      -0.3831
##      skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##      -0.9672      -0.3414      -0.4337       0.9741       0.0062
```

```r
shapiro.test(examData$lectures)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examData$lectures
## W = 1, p-value = 0.006
```




Normal Curve with Histogram
Number of Students
0.00
0.01
0.02
0.03
25
50
75
100
lectures
Number of Students
10.0
7.5
5.0
2.5
0.0



Impressions of the lecture results

The results here are widely distributed, and although there are a couple of peaks in the data, centered around 45% and 70%, it is generally evenly distributed across the range. There appears to be a somewhat normal distribution of the, as the graphs combining the histogram with the normal curve and probability plot show, but the number of students at the tails does not go to zero.

As with the previously evaluated variable, the numerical statistics confirm the visual impressions. In particular, the standard deviation of over 22 confirms the wide distribution of the results.

The Shapiro-Wilk test p-value of 0.006155 confirms that the lectures distribution is not normal.

Next let's look at the scores for the numeracy evaluation

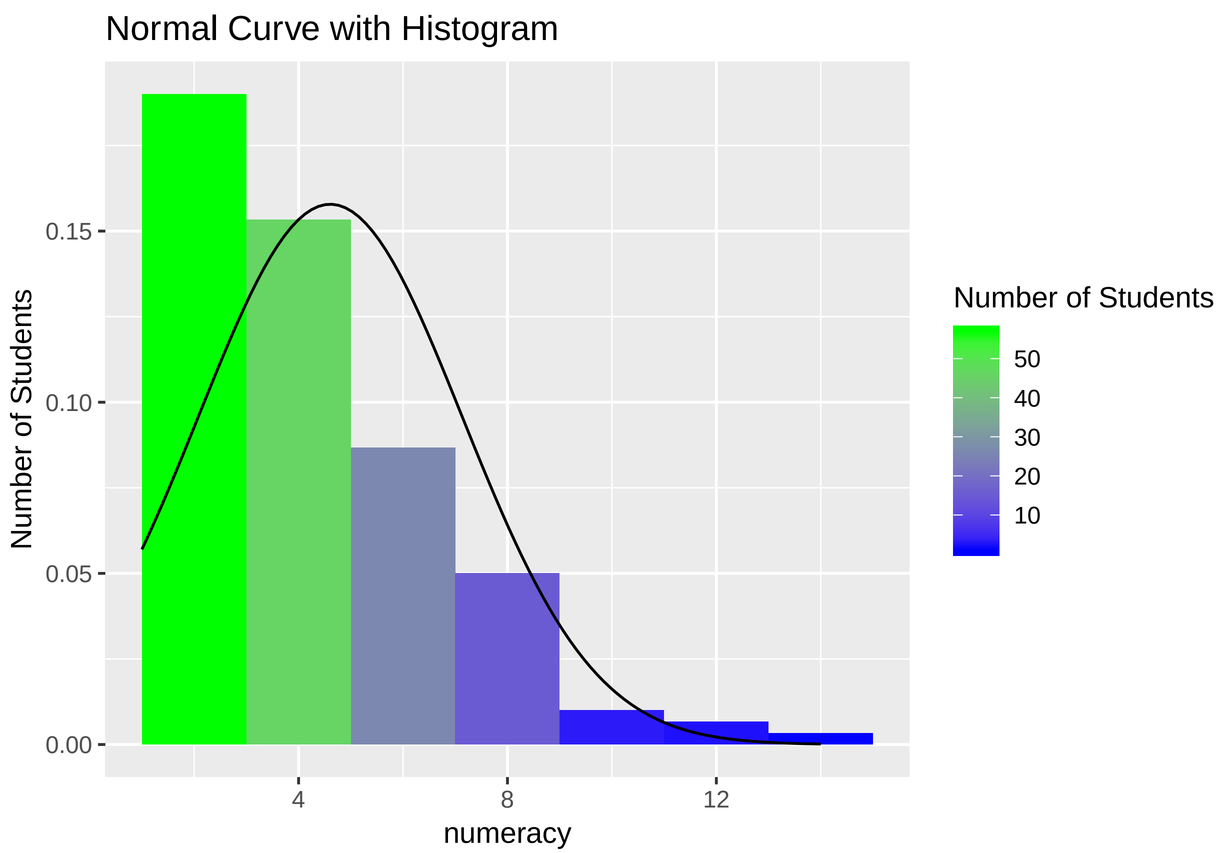
Remember to note that the scale here is 0-15.

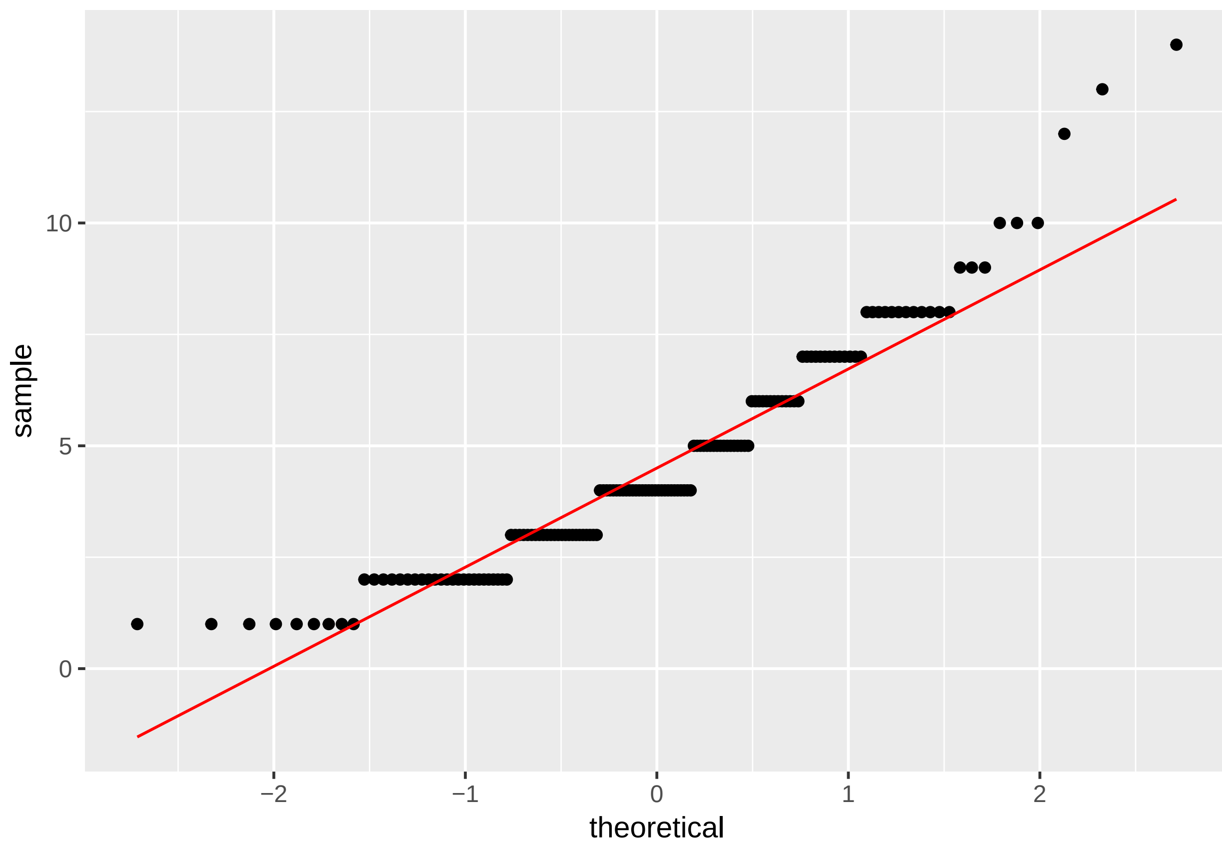
```
########################
numeracyStats <- stat.desc(examData$numeracy,norm = TRUE)
numeracyStats
```

```
##      nbr.val     nbr.null       nbr.na          min          max
## 150.00000000   0.00000000   0.00000000   1.00000000  14.00000000
##        range          sum       median         mean      SE.mean
##  13.00000000 691.00000000   4.00000000   4.60666667   0.20636335
## CI.mean.0.95          var      std.dev     coef.var     skewness
##   0.40777670   6.38787472   2.52742452   0.54864498   0.94116332
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##   2.37621457   0.96257582   1.22283687   0.92504009   0.00000046
```

```r
shapiro.test(examData$numeracy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examData$numeracy
## W = 0.9, p-value = 0.0000005
```


Normal Curve with Histogram
Number of Students
0.15
0.10
0.05
0.00
4
8
12
numeracy
Number of Students
50
40
30
20
10



Impressions of the numeracy results

Overall, assuming that higher scores are better, the students in general did poorly on this evaluation. Both the histograms and numerical results confirm, with the mean of 4.6 and median of 4 showing this conclusion. The data is highly skewed toward the lower end of the distribution, and there are only a few students who achieved a score greater than 10 (out of 15).

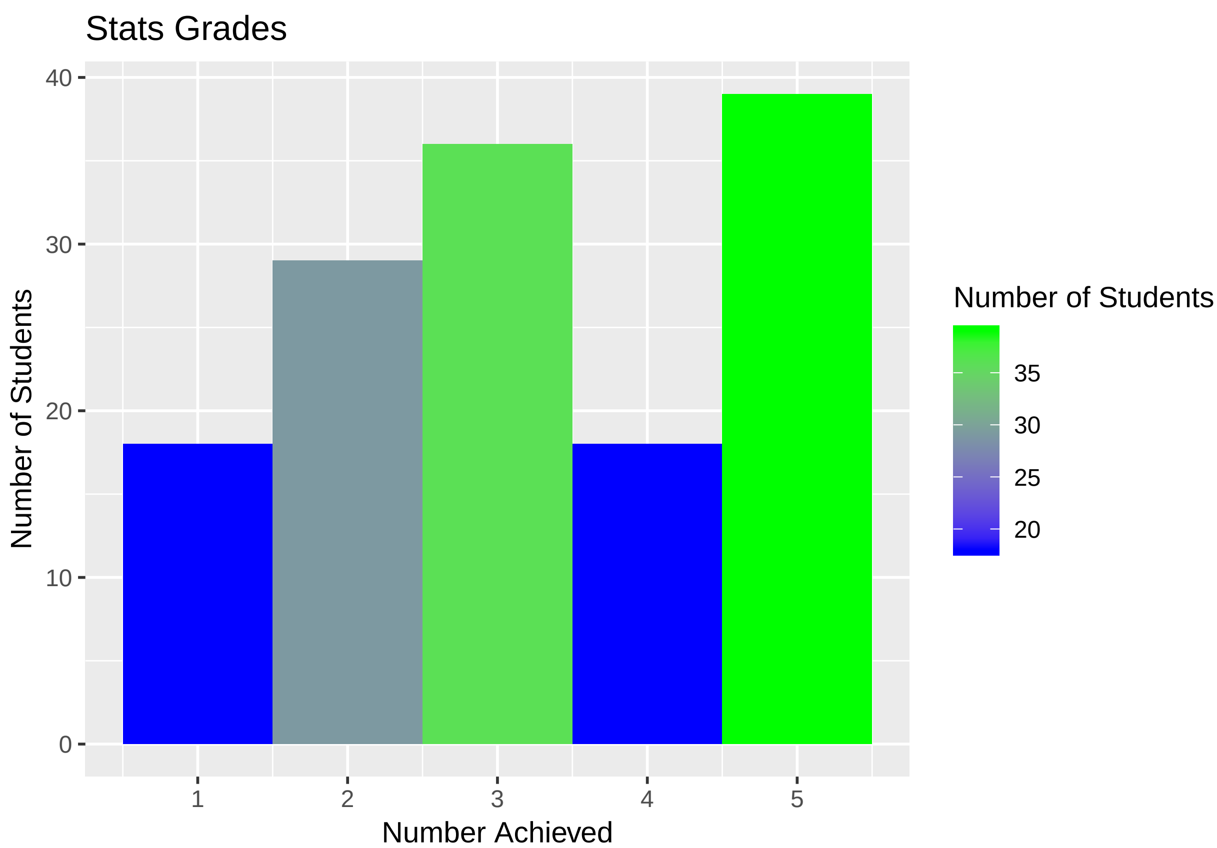
The Shapiro-Wilk test p-value of 0.0000004552 confirms that the exam distribution is not normal.

Finally, let's look at the letter grades for the students

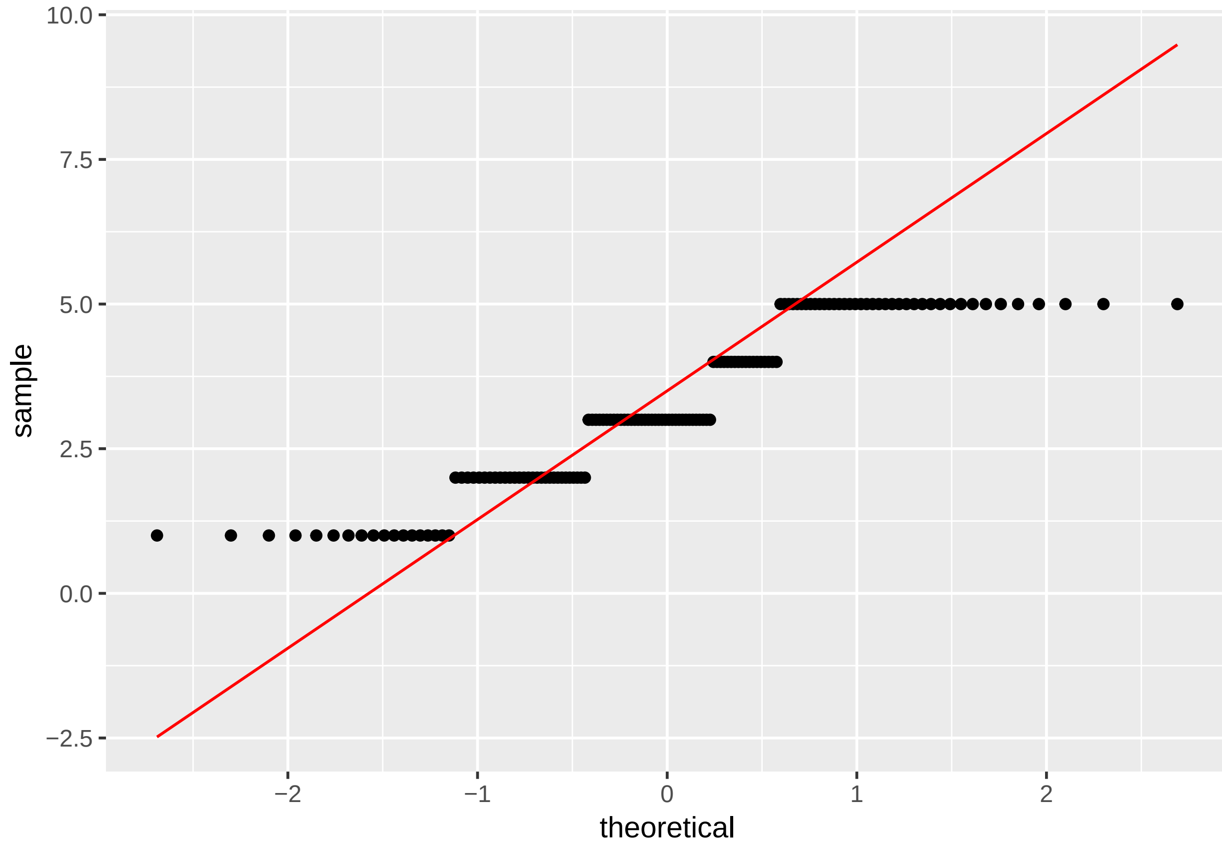
```
statsStats <- stat.desc(examData$stats,norm = TRUE)
statsStats
```

```
##         nbr.val        nbr.null          nbr.na             min             max
## 140.0000000000    0.0000000000   10.0000000000    1.0000000000    5.0000000000
##           range             sum          median            mean         SE.mean
##    4.0000000000  451.0000000000    3.0000000000    3.2214285714    0.1173831917
##    CI.mean.0.95             var         std.dev        coef.var        skewness
##    0.2320874318    1.9290339157    1.3888966541    0.4311430855   -0.0603549848
##        skew.2SE        kurtosis        kurt.2SE      normtest.W      normtest.p
##   -0.1473171214   -1.2727528476   -1.5638010103    0.8798778183    0.0000000029
```

```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```




```
## Warning: Removed 10 rows containing non-finite values (stat_qq).
## Warning: Removed 10 rows containing non-finite values (stat_qq_line).
```




```
##
##  Shapiro-Wilk normality test
##
## data:  examData$stats
## W = 0.9, p-value = 0.000000003
```

Impressions of the grade results

First, it is notable that 10 of the 150 students did not have grades recorded. Those records are excluded from the analysis.

Two things stand out about the distribution of grades.

First, a lot of students failed the class. For the professors and school administrators, this result might warrant a further examination of this group of students in a separate evaluation, to determine if there are underlying factors that would lead to this results.

Second, outside of the high number of failures, the grades distribution appears to be normally distributed around a "C" grade. The "average" grade is about a "C+", although only standard letter grades were issued.

The Shapiro-Wilk test p-value of 0.000000002869 confirms that the stats(student grades) distribution is not normal.

Is there a correlation between the evaluation results and the professor who taught the class?

Because each of the data distributions had multiple peaks, I am going to subset the student results by the professor who taught the class, to see if we can gain insights into whether the professor has an effect on student performance.

The methodology I will use is to first look at the student results graphically, to see if I can detect any patterns in the data, then I will follow up with statistics, to further analyze the student results.

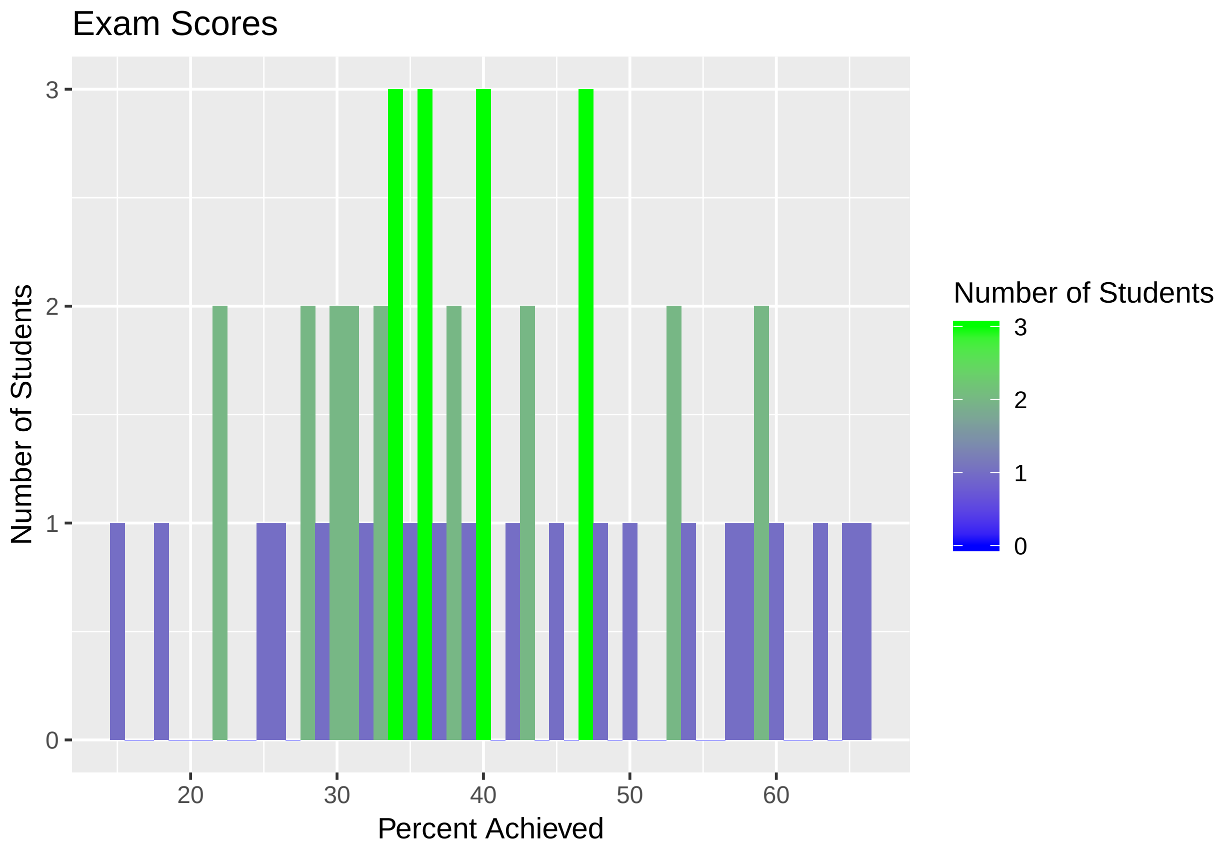
I'm going to take a different approach with the individual professor groupings, from what I did above. Because I want to see if the professor teaching has an effect on the results achieved, I'm going to look at all of the histograms as a group, to see if there is on overall pattern to the results. From there, I will then examine them individually, to see what we can determine about each of the evaluations. Finally, I will examine the numerical statistics, to see what they tell us.

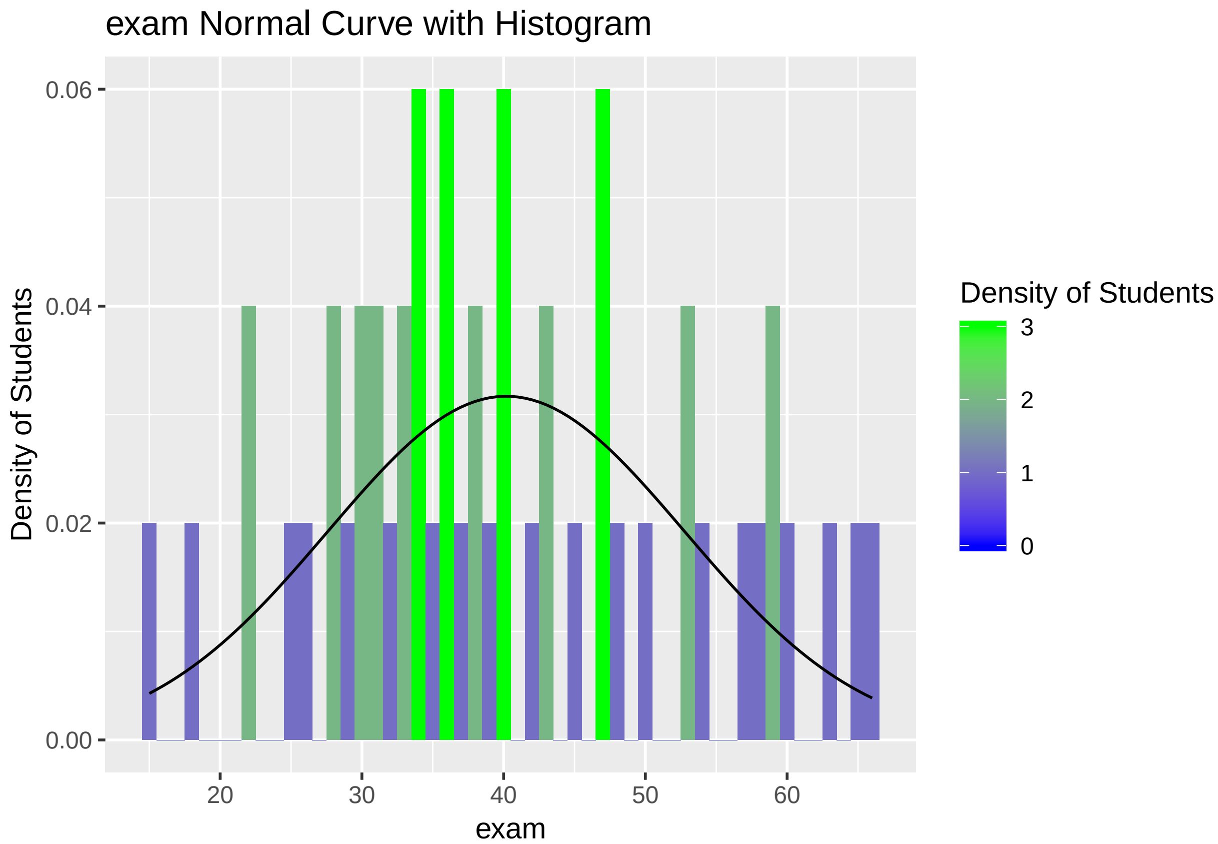
One change that I will make in how I look at the data is that I will change my data groupings (in the histograms) from a range of 2 to a range of 1. In my graphs above, I have set the binwidth to 2, as a compromise on how much granularity to show in the data. So my percentage results are in groupings of 2%, to help with easier data visualization. Because there will be fewer students represented within the professor subsets, I will set the binwidth to 1%, and allow the data to be more granular, to attempt to better see patterns.

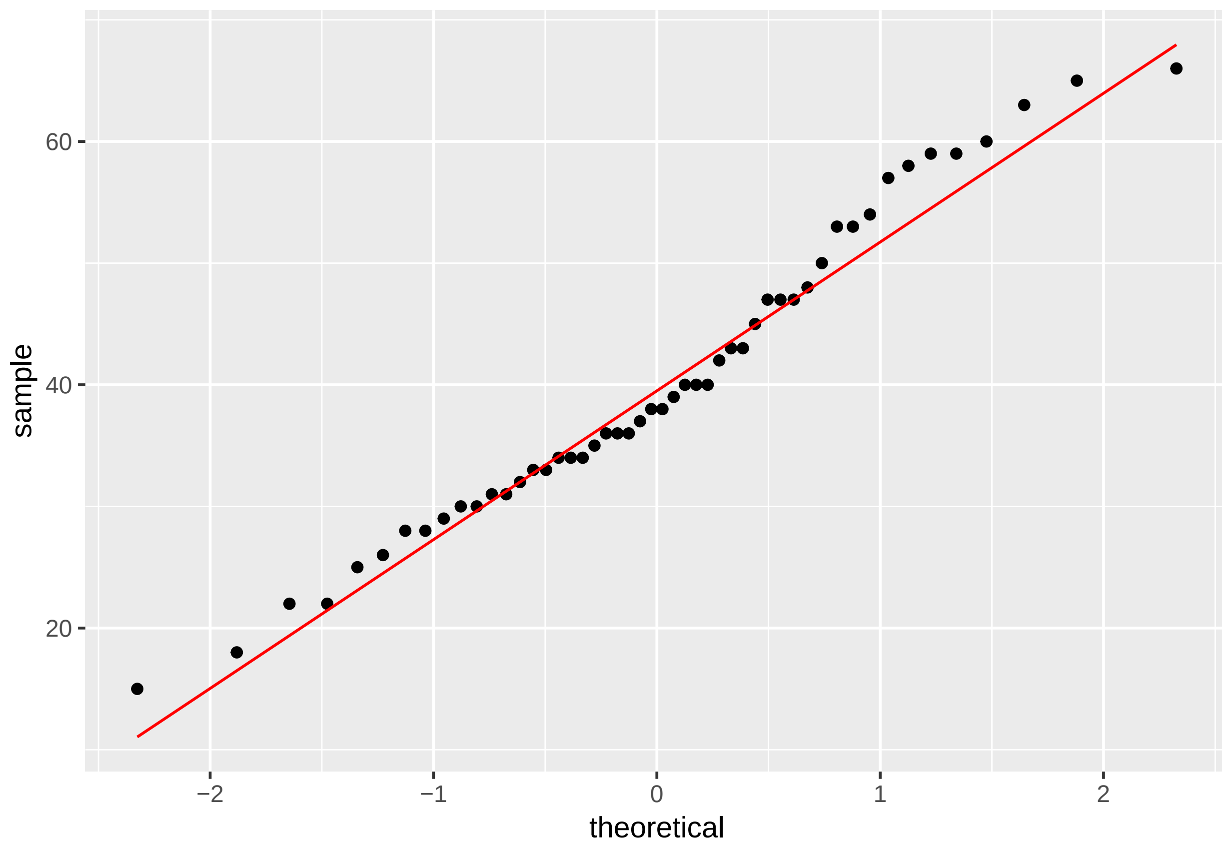
First, we will look at the student results for Professor George.

```
#subset the data for Professor George
examDataGeorge <- subset(examData, uni=="0",
                         select=c("exam", "computer", "lectures", "numeracy", "stats"))
str(examDataGeorge)
```

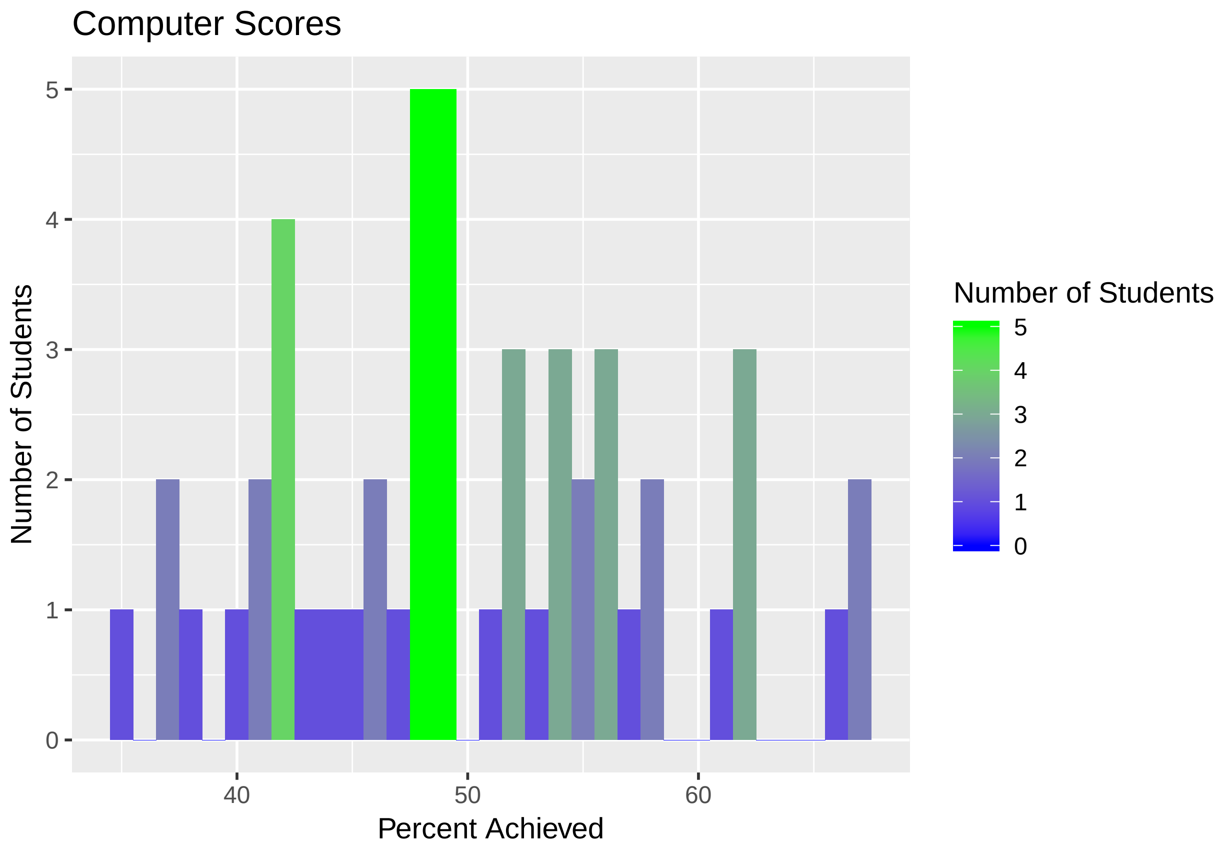
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    50 obs. of  5 variables:
##  $ exam    : num  18 30 40 30 40 15 36 40 63 31 ...
##  $ computer: num  54 47 58 37 53 48 49 49 45 62 ...
##  $ lectures: num  75 8.5 69.5 67 44.5 76.5 70 18.5 43.5 100 ...
##  $ numeracy: num  7 1 6 6 2 8 3 7 4 6 ...
##  $ stats   : num  1 1 1 2 2 2 2 3 3 3 ...
```

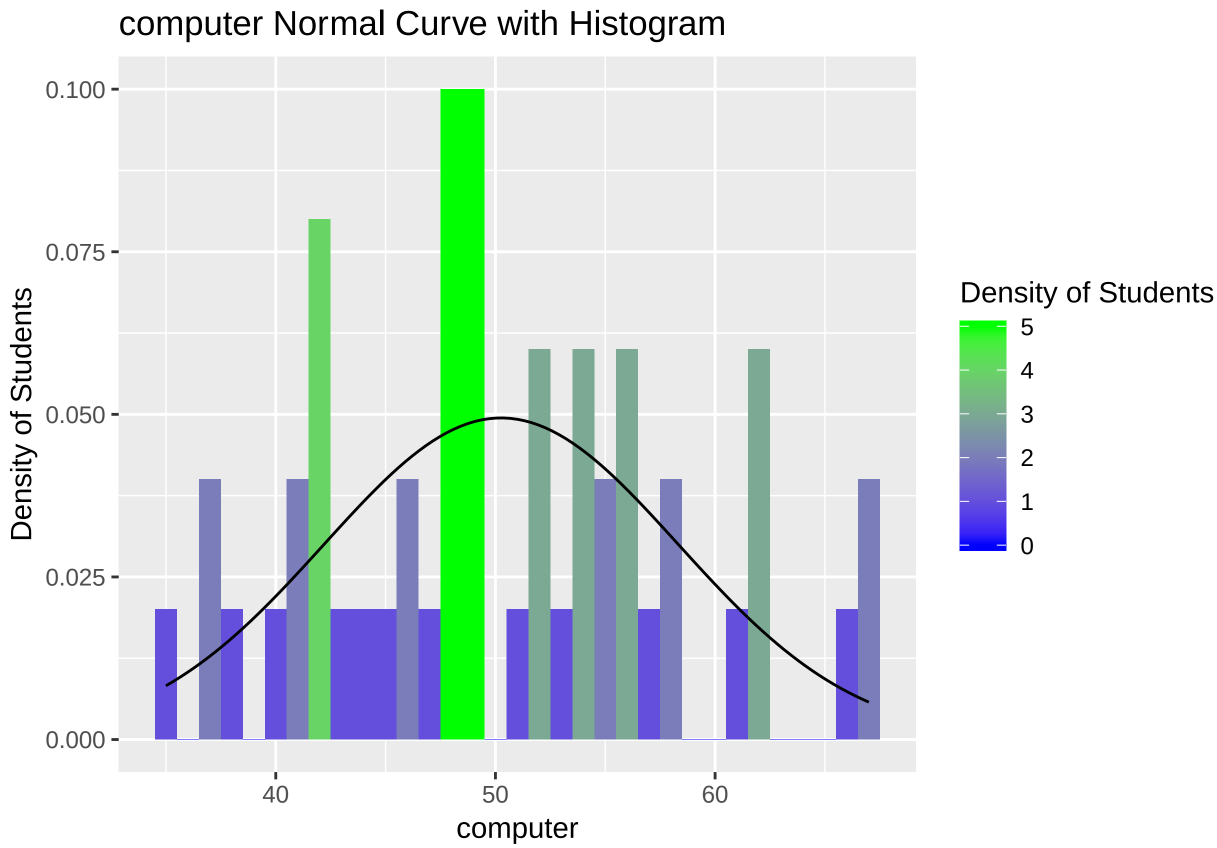

Exam Scores
Number of Students
Percent Achieved
0
1
2
3
20
30
40
50
60
Number of Students
3
2
1
0

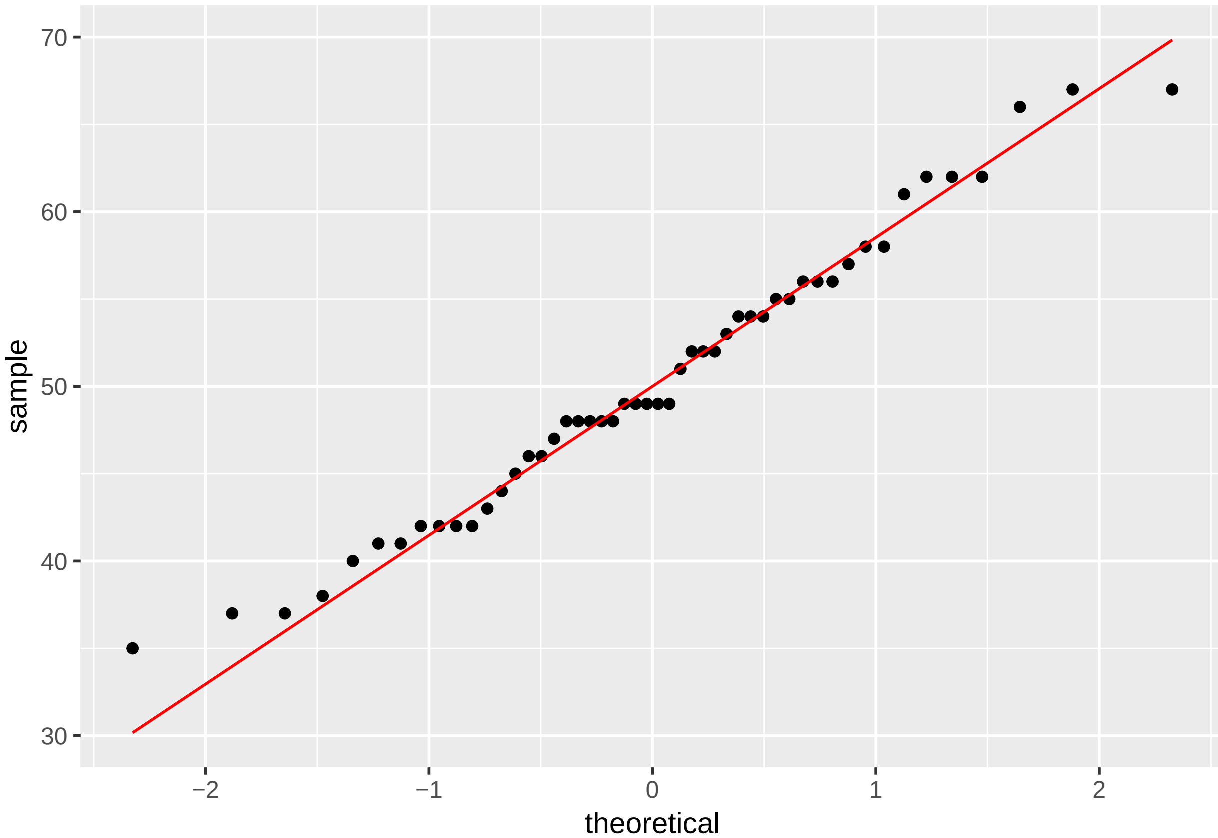

exam Normal Curve with Histogram
0.06
0.04
0.02
0.00
Density of Students
20
30
40
50
60
exam
Density of Students
3
2
1
0



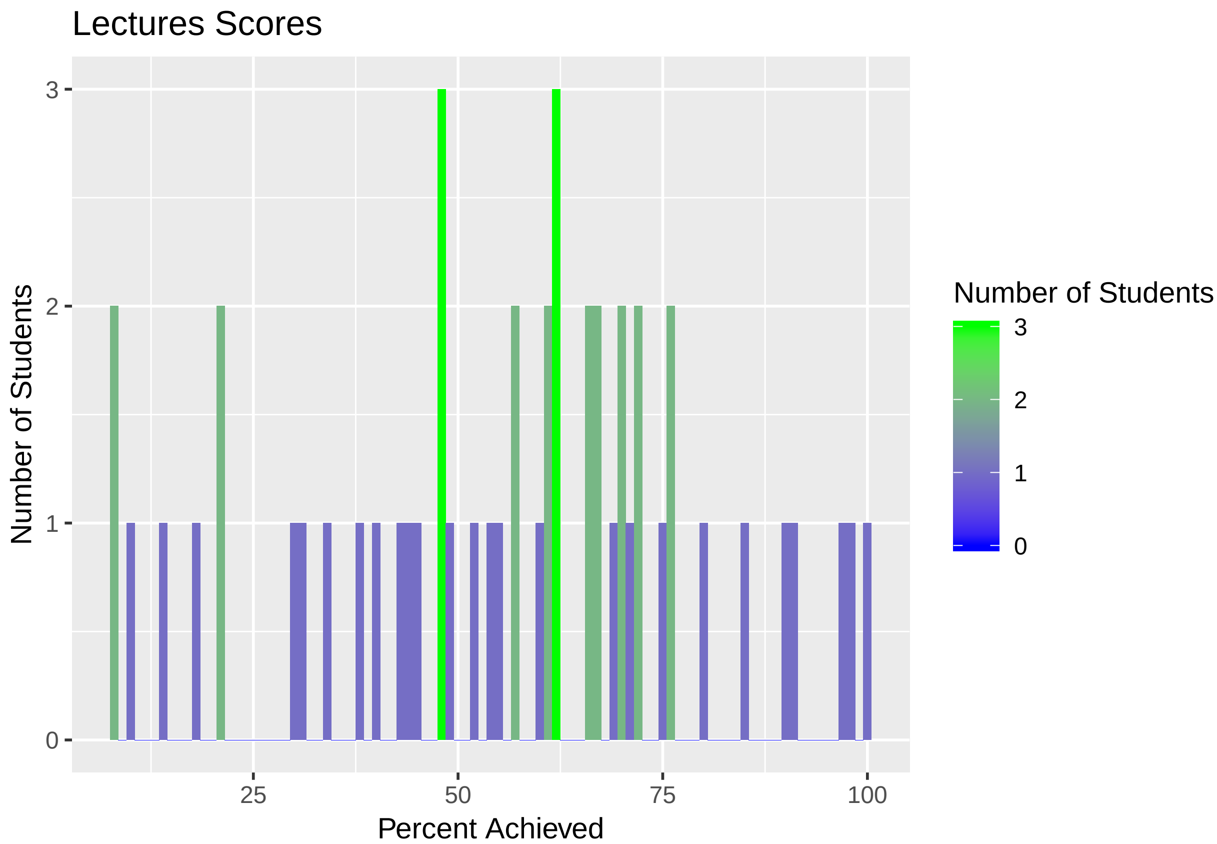
60
40
20
sample
−2
−1
0
1
2
theoretical

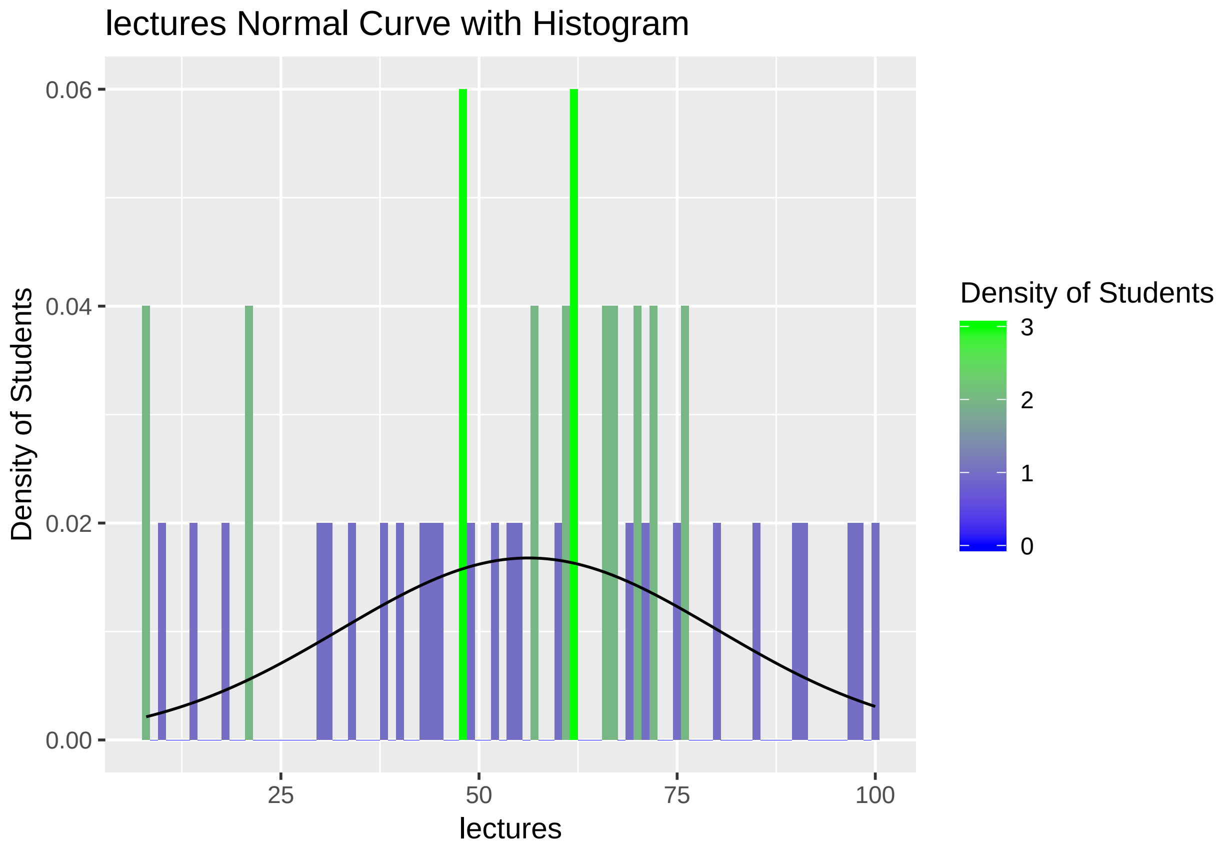

Computer Scores
Number of Students
Percent Achieved
0
1
2
3
4
5
40
50
60
Number of Students
5
4
3
2
1
0


computer Normal Curve with Histogram
0.100
0.075
0.050
0.025
0.000
Density of Students
40
50
60
computer
Density of Students
5
4
3
2
1
0

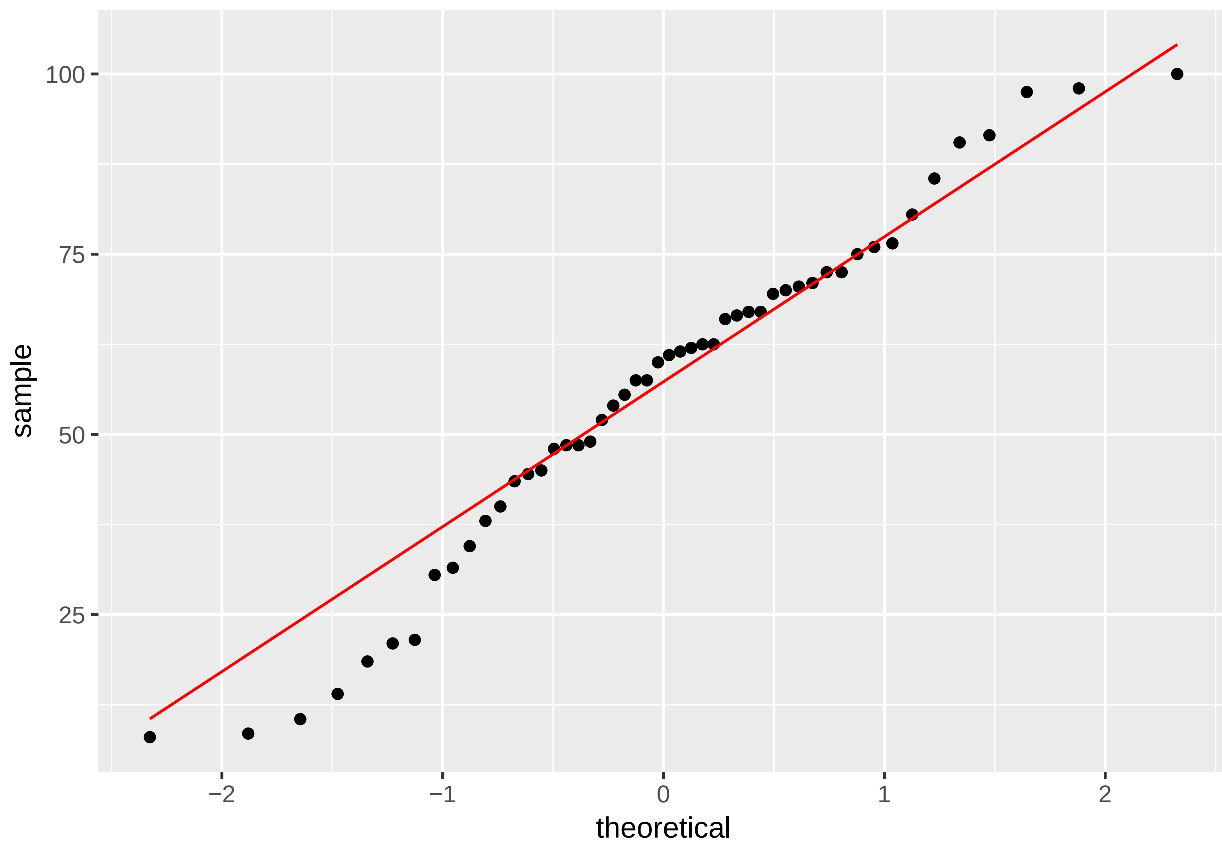


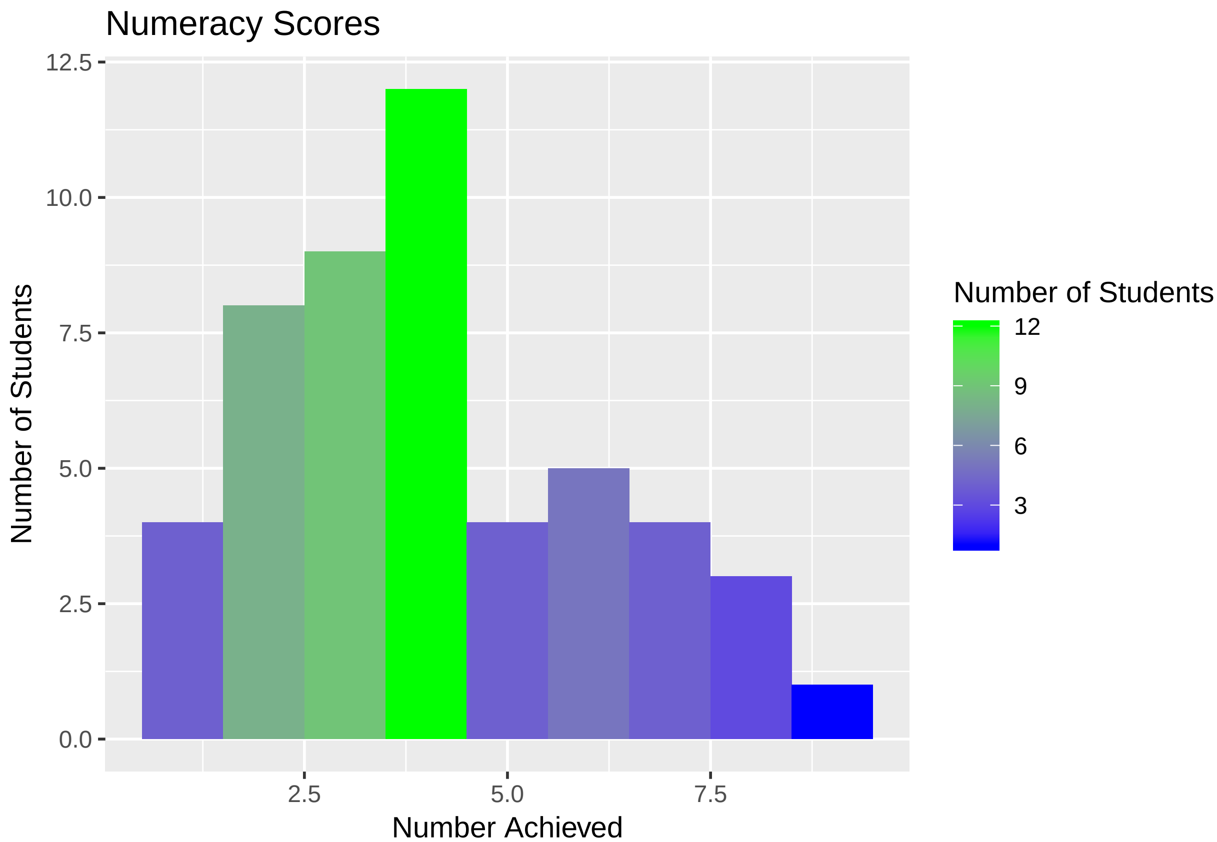
70
60
50
40
30
sample
−2
−1
0
1
2
theoretical

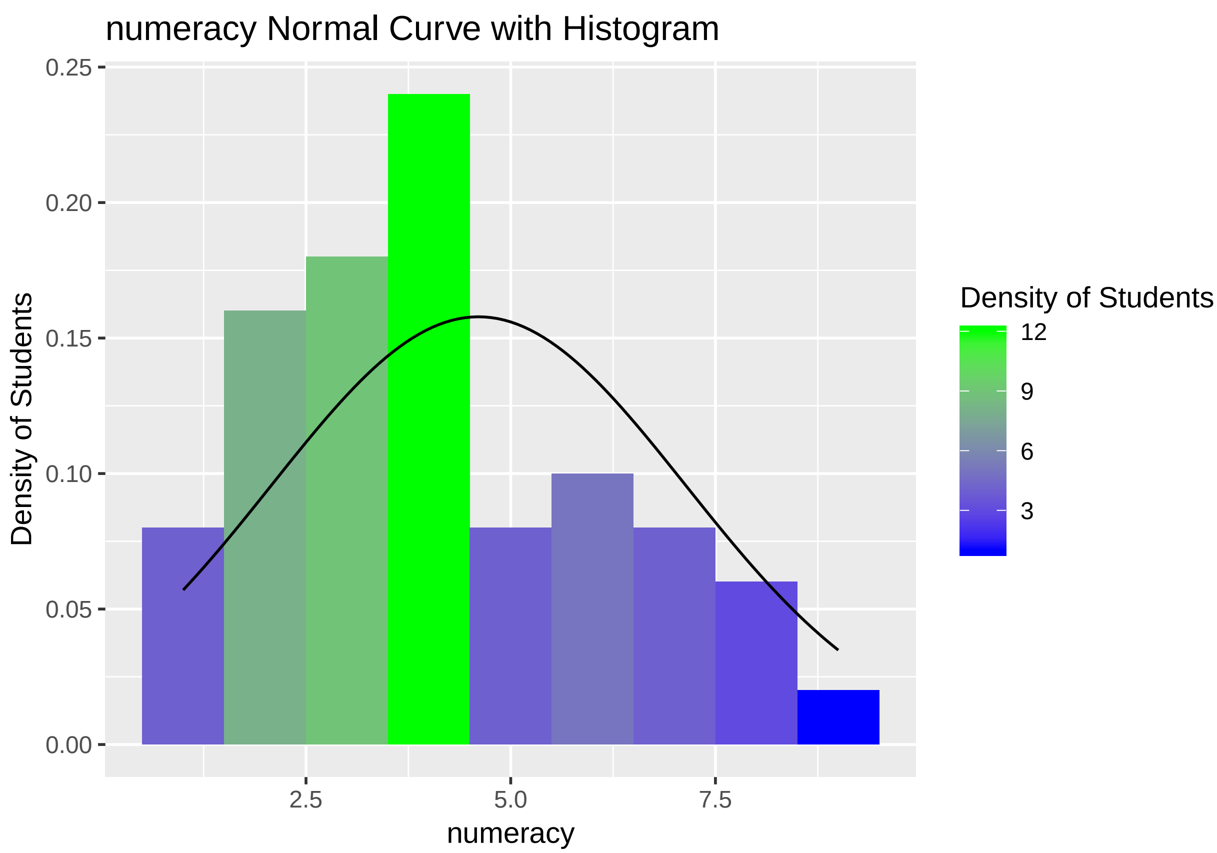

Lectures Scores
Number of Students
Percent Achieved
0
1
2
3
25
50
75
100
Number of Students
3
2
1
0

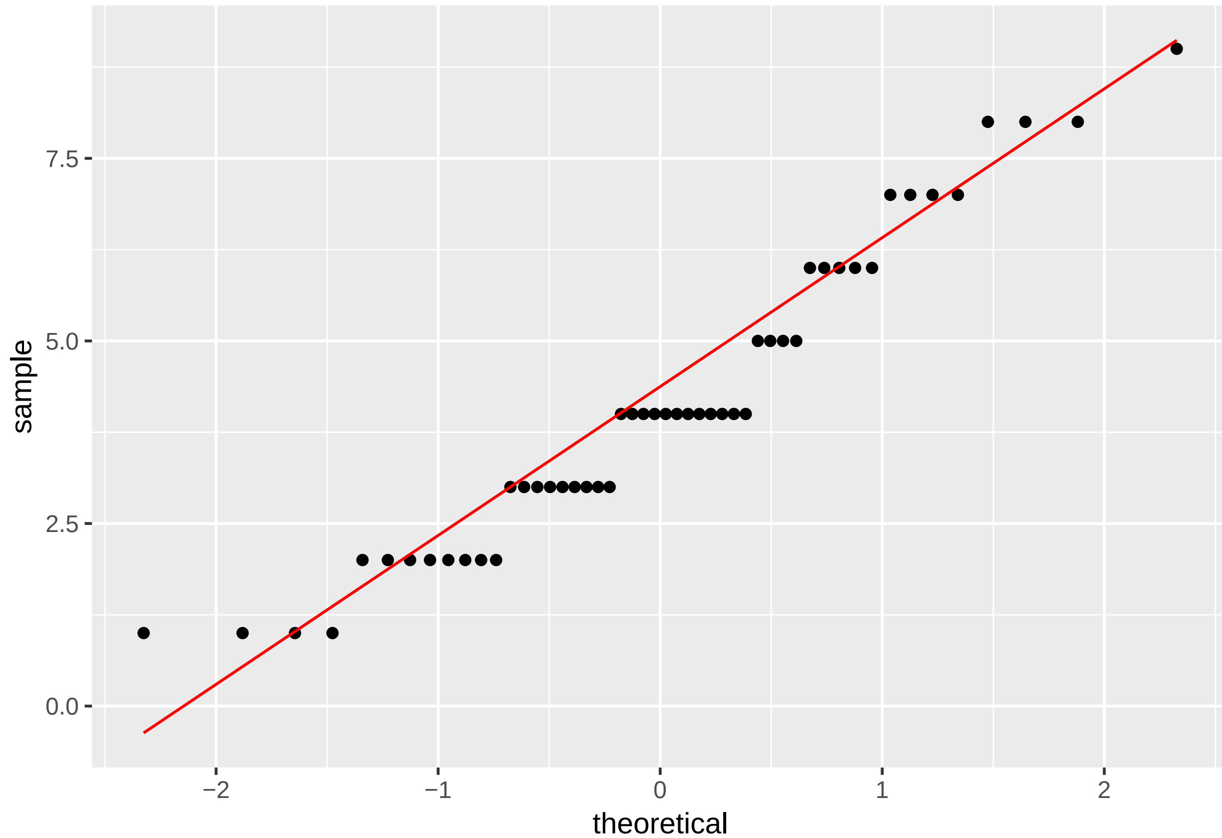


lectures Normal Curve with Histogram
0.06
0.04
0.02
0.00
Density of Students
25
50
75
100
lectures
Density of Students
3
2
1
0

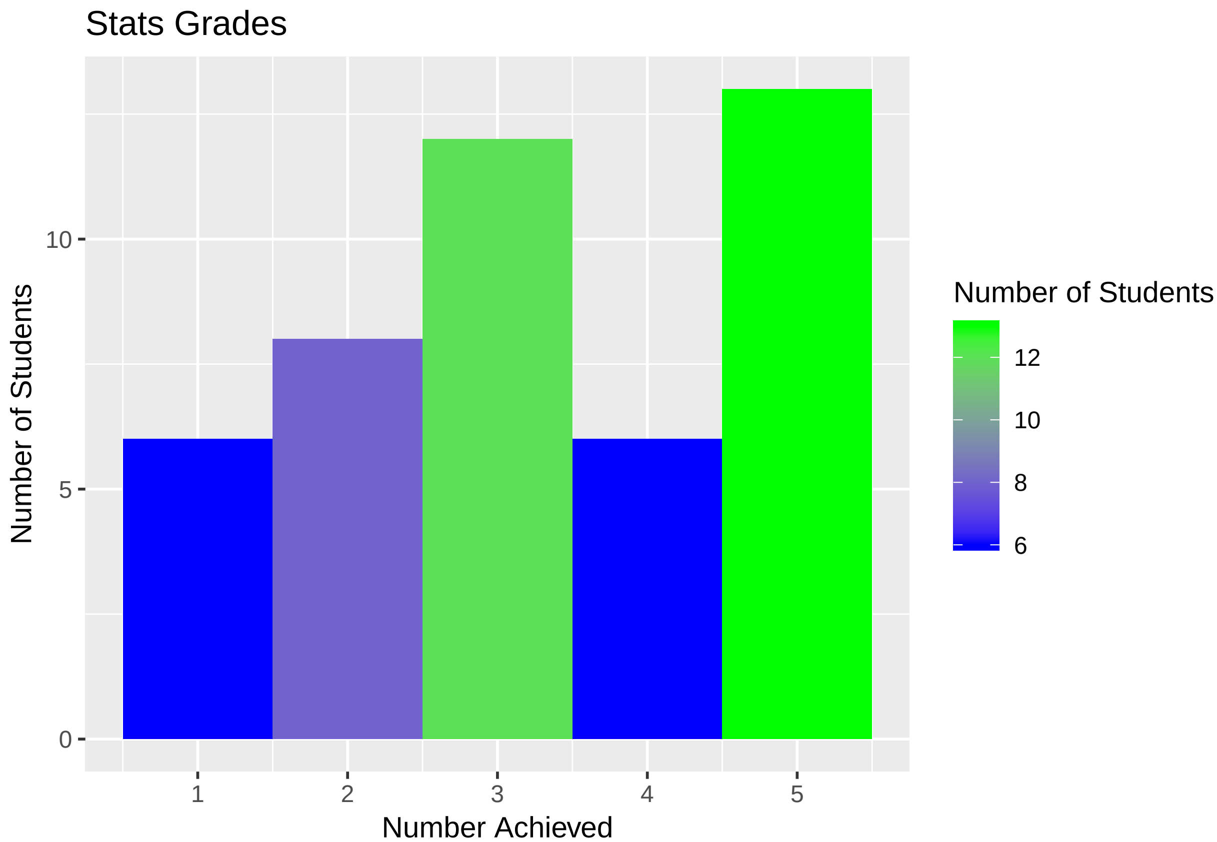

100
75
50
25
sample
−2
−1
0
1
2
theoretical


Numeracy Scores
Number of Students
Number Achieved
12.5
10.0
7.5
5.0
2.5
0.0
2.5
5.0
7.5
Number of Students
12
9
6
3

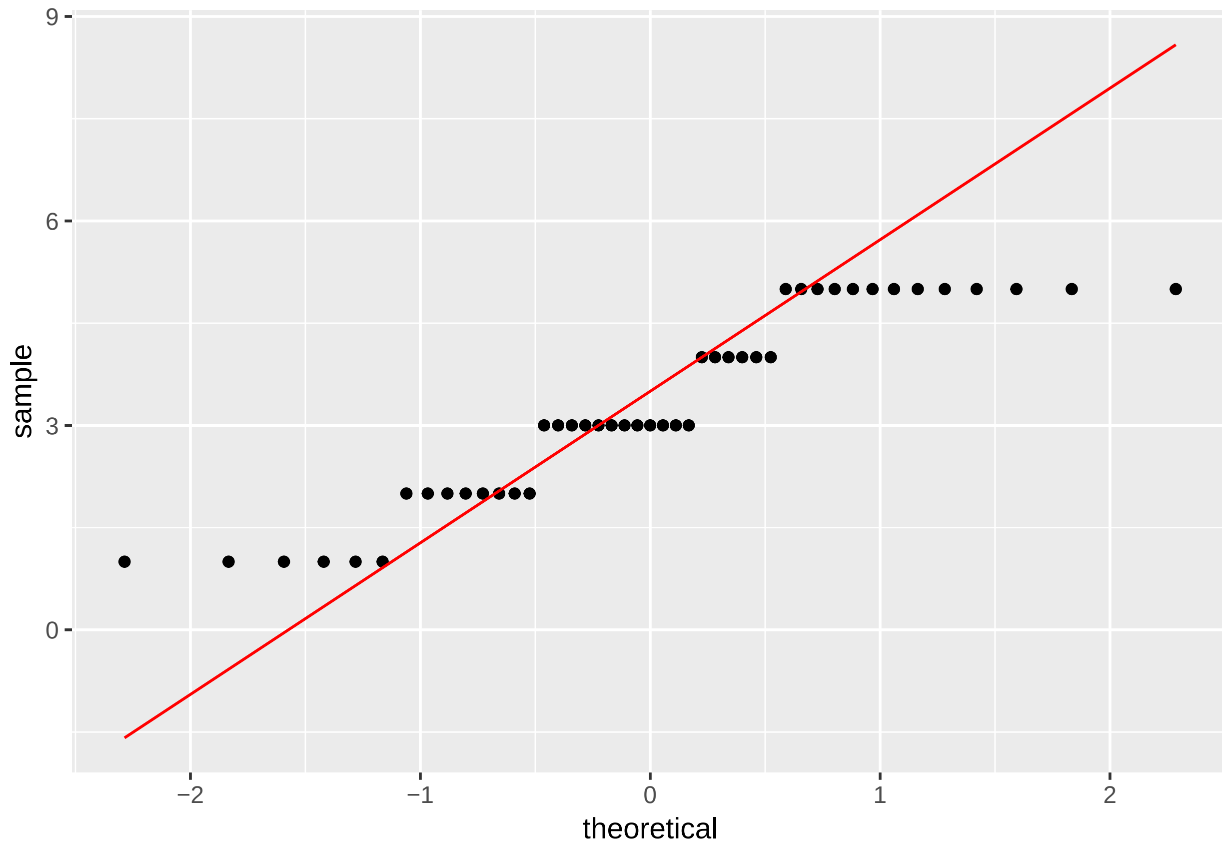

numeracy Normal Curve with Histogram
0.25
0.20
0.15
0.10
0.05
0.00
Density of Students
2.5
5.0
7.5
numeracy
Density of Students
12
9
6
3




```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```




```
## Warning: Removed 5 rows containing non-finite values (stat_qq).
## Warning: Removed 5 rows containing non-finite values (stat_qq_line).
```




```
#run statistics for each of the evaluations
examStatsGeorge     <- stat.desc(examDataGeorge$exam, norm = TRUE)
computerStatsGeorge <- stat.desc(examDataGeorge$computer, norm = TRUE)
lecturesStatsGeorge <- stat.desc(examDataGeorge$lectures, norm = TRUE)
numeracyStatsGeorge <- stat.desc(examDataGeorge$numeracy,norm = TRUE)
statsStatsGeorge    <- stat.desc(examDataGeorge$stats,norm = TRUE)

examStatsGeorge
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##        50.00         0.00         0.00        15.00        66.00
##        range          sum       median         mean      SE.mean
##        51.00      2009.00        38.00        40.18         1.78
## CI.mean.0.95          var      std.dev     coef.var     skewness
##         3.58       158.48        12.59         0.31         0.29
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##         0.43        -0.72        -0.55         0.97         0.28
```

```
computerStatsGeorge
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##        50.00         0.00         0.00        35.00        67.00
##        range          sum       median         mean      SE.mean
##        32.00      2513.00        49.00        50.26         1.14
## CI.mean.0.95          var      std.dev     coef.var     skewness
##         2.29        65.09         8.07         0.16         0.21
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
```

```
##          0.32        -0.68        -0.51         0.98         0.46
```

```r
lecturesStatsGeorge
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##        50.00         0.00         0.00         8.00       100.00
##        range          sum       median         mean      SE.mean
##        92.00      2813.00        60.50        56.26         3.36
## CI.mean.0.95          var      std.dev     coef.var     skewness
##         6.76       565.14        23.77         0.42        -0.29
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##        -0.43        -0.56        -0.43         0.97         0.23
```

```r
numeracyStatsGeorge
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##       50.000        0.000        0.000        1.000        9.000
##        range          sum       median         mean      SE.mean
##        8.000      206.000        4.000        4.120        0.292
## CI.mean.0.95          var      std.dev     coef.var     skewness
##        0.587        4.271        2.067        0.502        0.482
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##        0.715       -0.652       -0.492        0.941        0.015
```

```r
statsStatsGeorge
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##     45.00000      0.00000      5.00000      1.00000      5.00000
##        range          sum       median         mean      SE.mean
##      4.00000    147.00000      3.00000      3.26667      0.20938
## CI.mean.0.95          var      std.dev     coef.var     skewness
##      0.42197      1.97273      1.40454      0.42996     -0.13068
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##     -0.18471     -1.29768     -0.93420      0.87938      0.00023
```

```r
shapiro.test(examDataGeorge$exam)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataGeorge$exam
## W = 1, p-value = 0.3
```

```r
shapiro.test(examDataGeorge$computer)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataGeorge$computer
## W = 1, p-value = 0.5
```

```r
shapiro.test(examDataGeorge$lectures)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataGeorge$lectures
## W = 1, p-value = 0.2
```

```r
shapiro.test(examDataGeorge$numeracy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataGeorge$numeracy
## W = 0.9, p-value = 0.01
```

```r
shapiro.test(examDataGeorge$stats)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataGeorge$stats
## W = 0.9, p-value = 0.0002
```

What do the Professor George results show?

First, a bit about the students themselves. 1. 50 of the students in the cohort of 150 were in class with Professor George. 2. Of those 50, 5 students do not have grades recorded (stats variable).

The first thing that stands out for the Professor George cohort is that the numerical evaluation scores show a much more normal distribution within the cohort than for the overall student population. What this tells us is that the fact that Professor George is teaching the class has an effect on student performance. And being brutally honest, it appears that Professor George's teaching method is not a good enabler of student success. There are two results from the visuals that lead me to this conclusion:

1. The three percentage-based histograms all center the normal curve around 50%, either slightly above or below. The scores are not tightly grouped, with long and significant tails, but they do all show a somewhat normal distribution.
2. The numeracy scores center around 4-5, with the highest result only be 8, out of 15. The numeracy scores are skewed to the low end of the range, indicating that the students in general did not do well on this evaluation.

Additionally, looking at the final student grades within the Professor George cohort, there are two conclusions from the histogram.

1. The grades distribution for this cohort looks very similar to that of the overall student population, leading us to initially conclude that the final measure of student success is not positively or negatively affected by Professor George's being the instructor.
2. A high number of student (13 of 45 with grades) failed the class.

Finally, looking at the statistics confirms the conclusions that I reached from the visual representations.

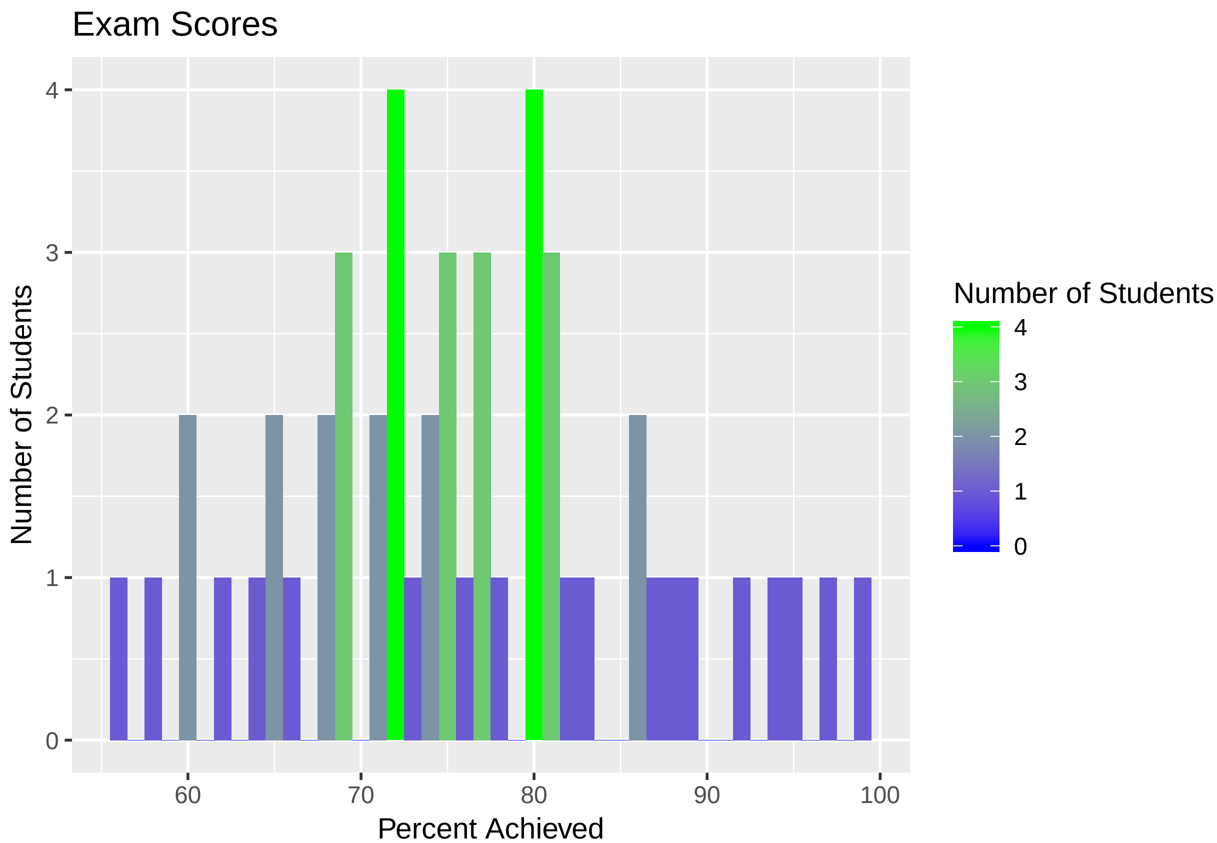
1. The means and medians confirm that the students performed less than admirably in this cohort, while the standard deviations confirm the spread of the results.
2. The skew and kurtosis values again confirm the results seen in the histograms.
3. The Shapiro-Wilk tests for exam, computer, and lectures all confirm that their scores are all normally distributed, while the same tests for numeracy and stats confirm that those scores are not normally distributed. Again, this confirms what we are seeing graphically.

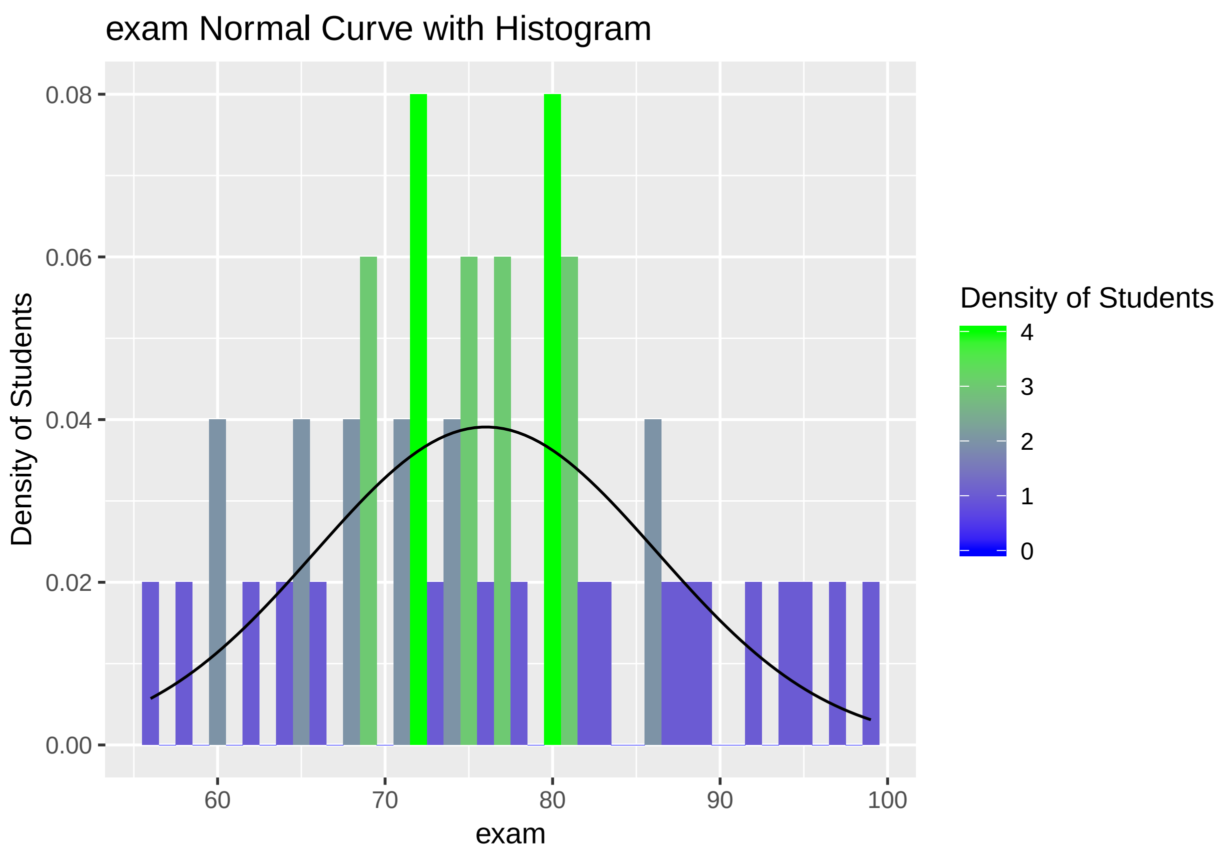
Now let's take a look at Professor Jeff's results.

```r
#subset the data for Professor Jeff
examDataJeff  <- subset(examData, uni=="1",
```

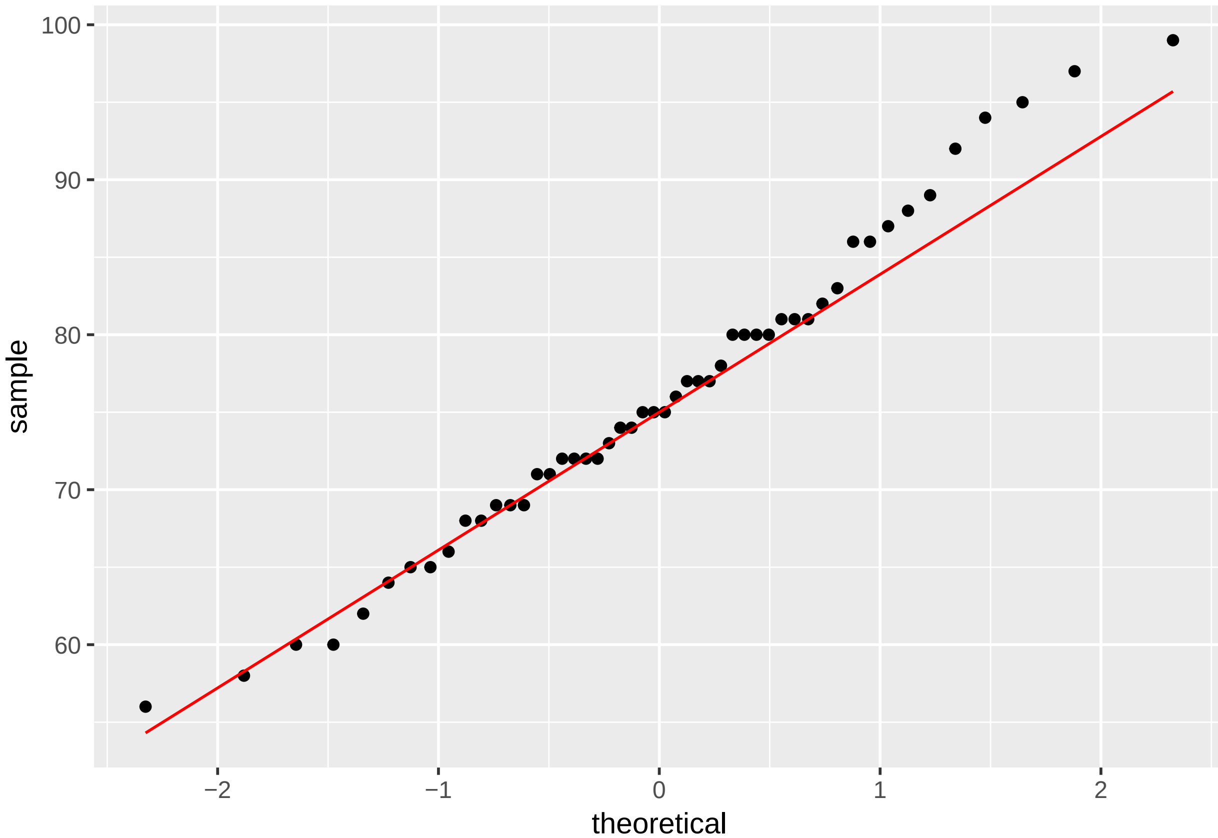
```r
                              select=c("exam", "computer", "lectures", "numeracy", "stats"))
str(examDataJeff)
```

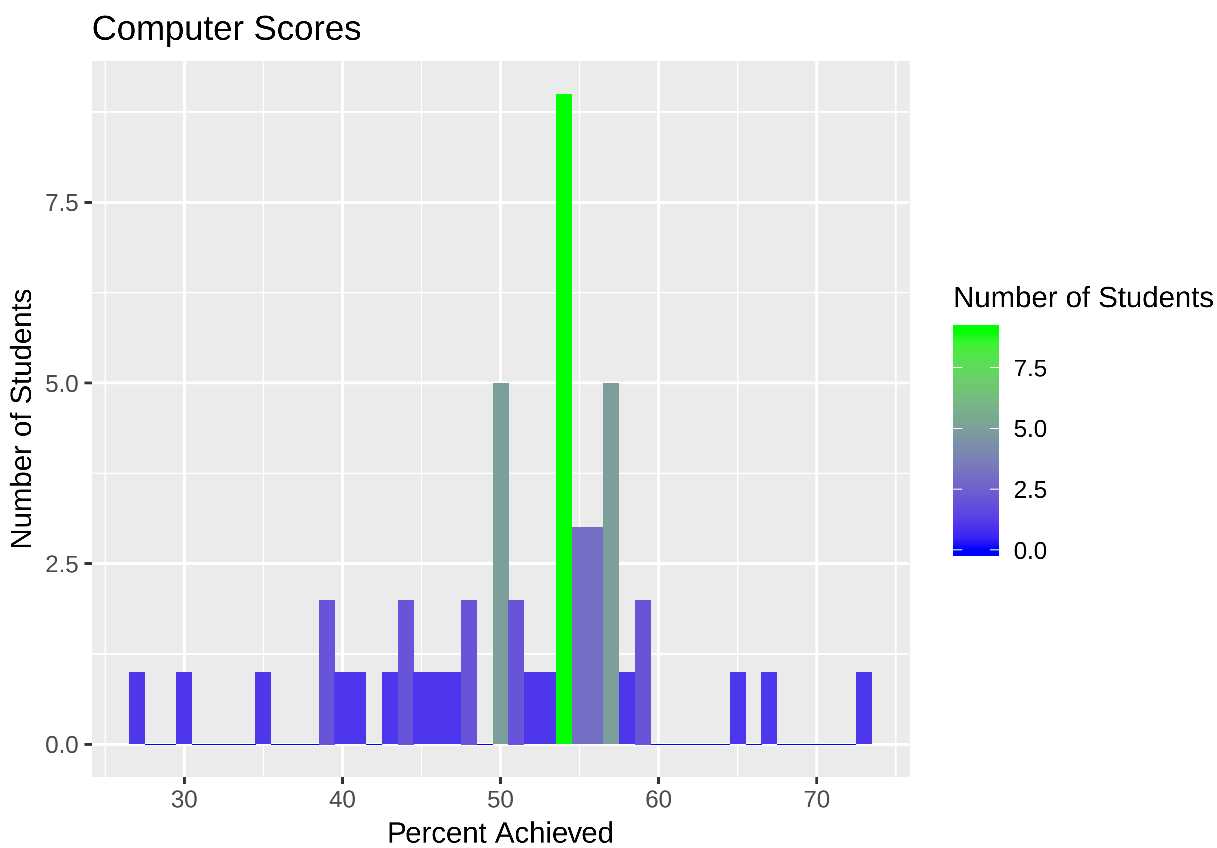
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    50 obs. of  5 variables:
##  $ exam    : num  56 76 72 77 77 66 62 86 97 72 ...
##  $ computer: num  30 48 54 44 54 58 59 54 35 56 ...
##  $ lectures: num  84.5 51 58.5 42 65.5 56 71.5 48.5 84.5 47.5 ...
##  $ numeracy: num  7 8 5 6 9 7 2 5 5 2 ...
##  $ stats   : num  1 1 1 2 2 2 2 3 3 3 ...
```

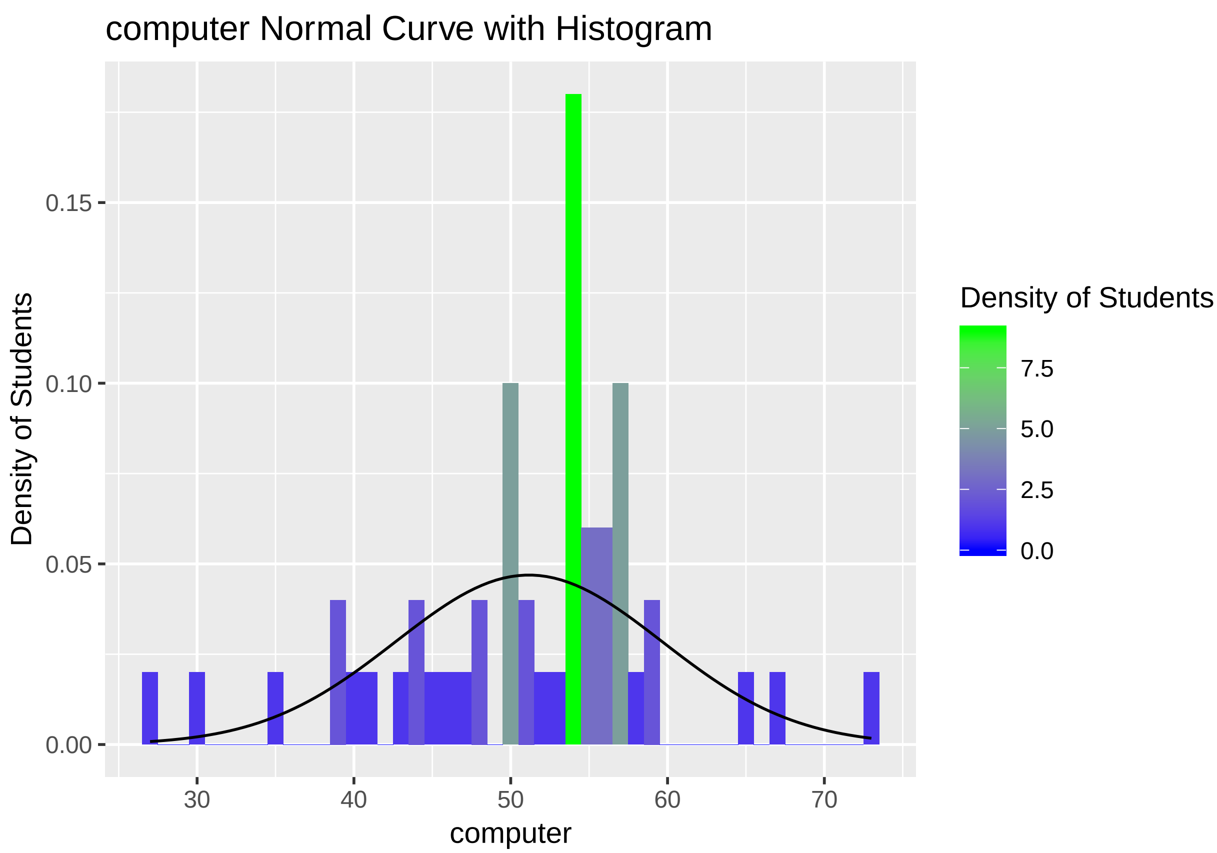


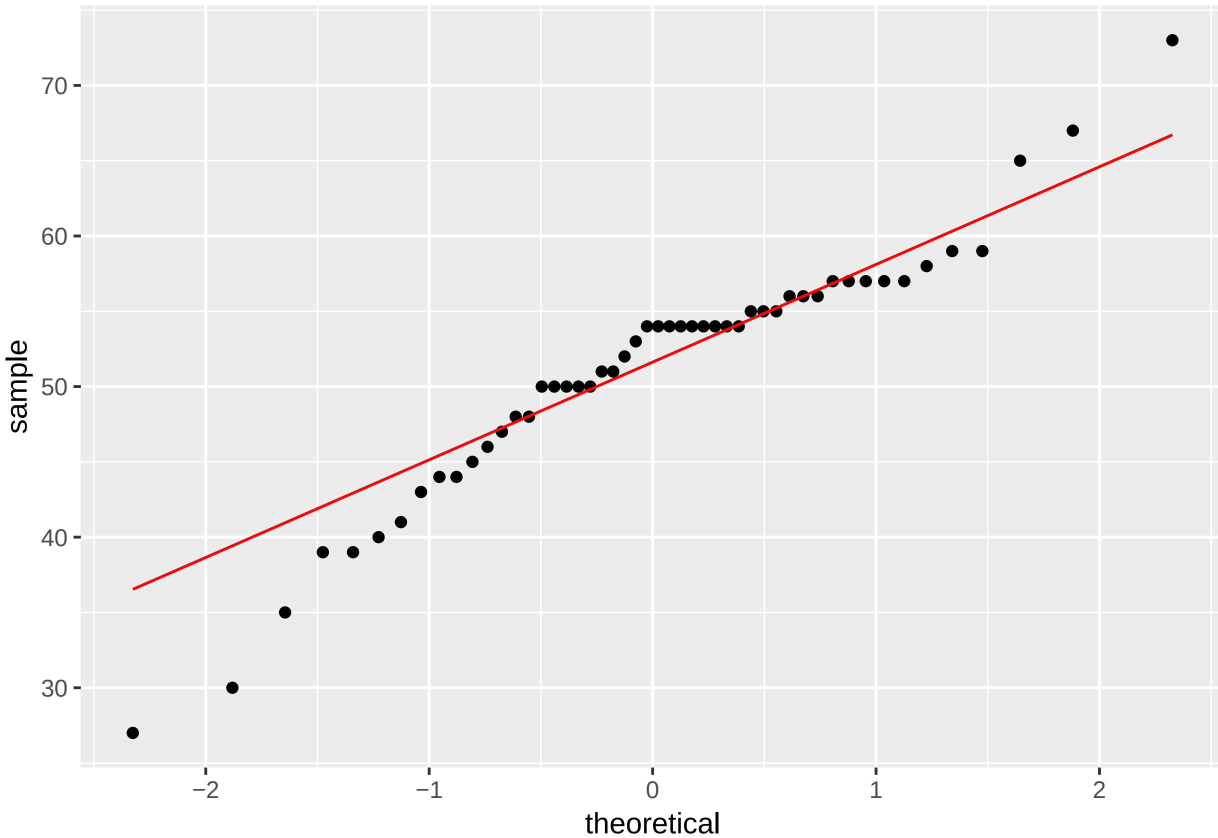


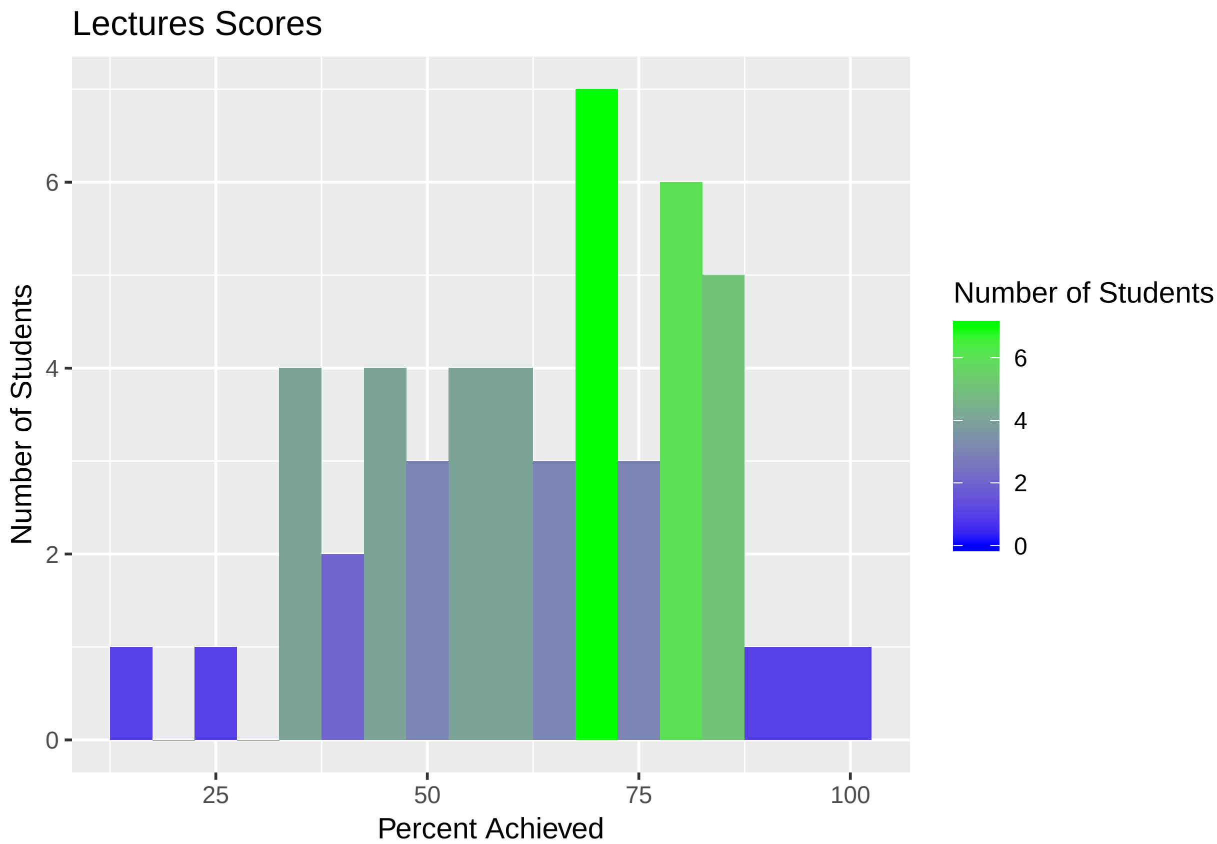
exam Normal Curve with Histogram
Density of Students
0.08
0.06
0.04
0.02
0.00
60
70
80
90
100
exam
Density of Students
4
3
2
1
0

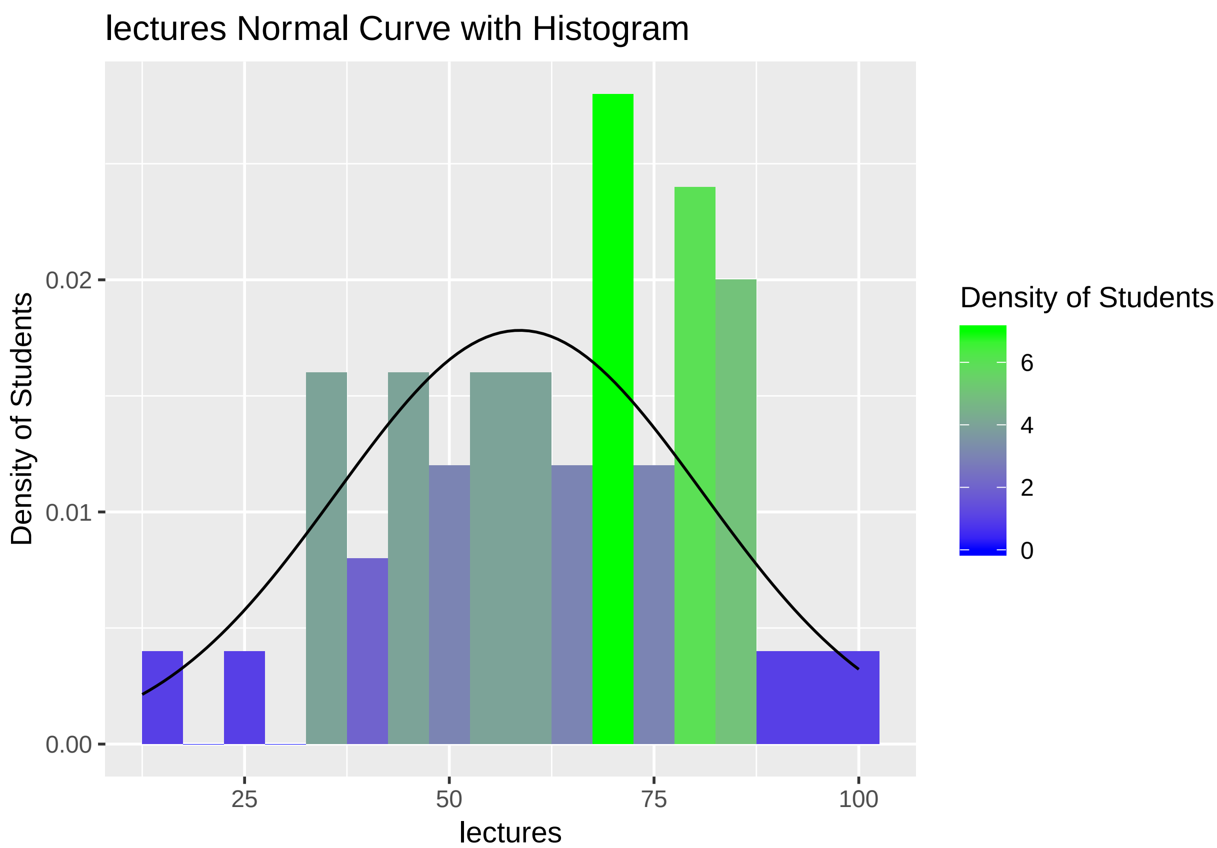

100
90
80
70
60
sample
−2
−1
0
1
2
theoretical

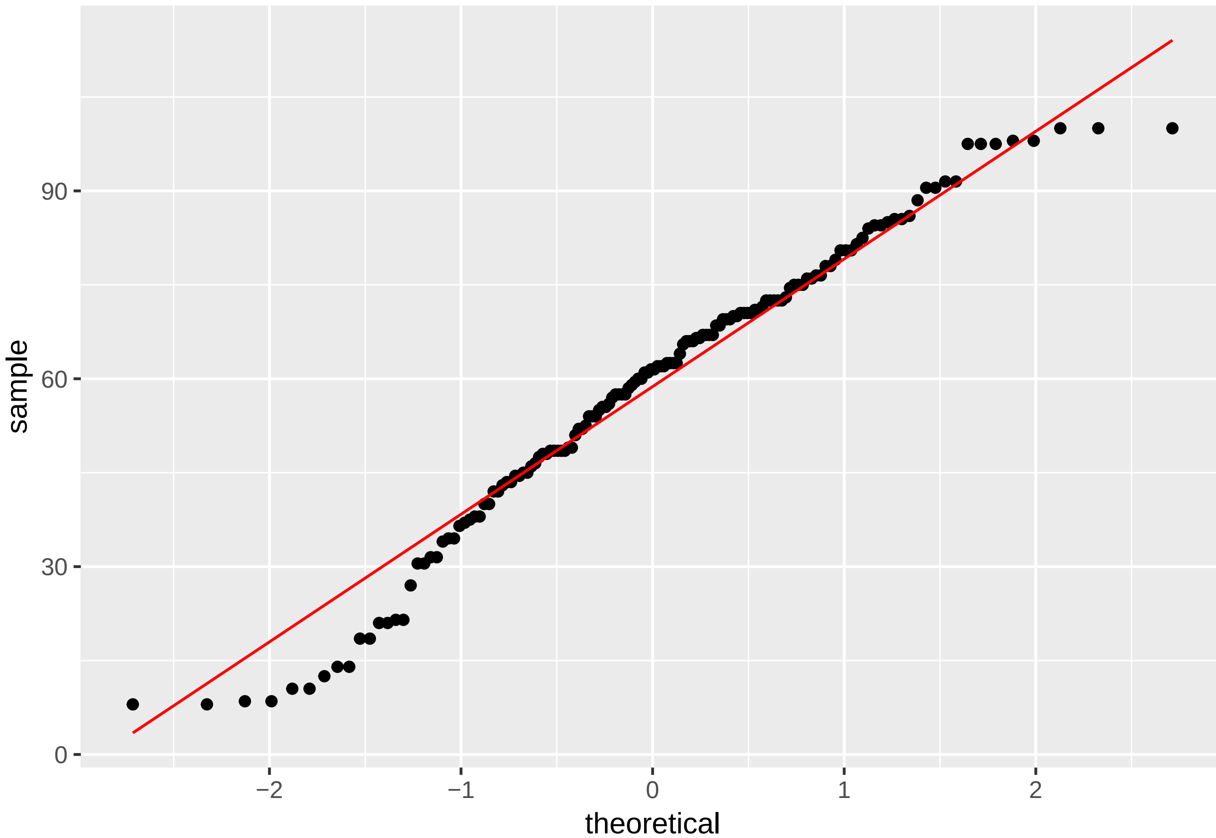

Computer Scores
Number of Students
7.5
5.0
2.5
0.0
30
40
50
60
70
Percent Achieved
Number of Students
7.5
5.0
2.5
0.0


computer Normal Curve with Histogram
Density of Students
0.15
0.10
0.05
0.00
30
40
50
60
70
computer
Density of Students
7.5
5.0
2.5
0.0


70
60
50
40
30
sample
−2
−1
0
1
2
theoretical

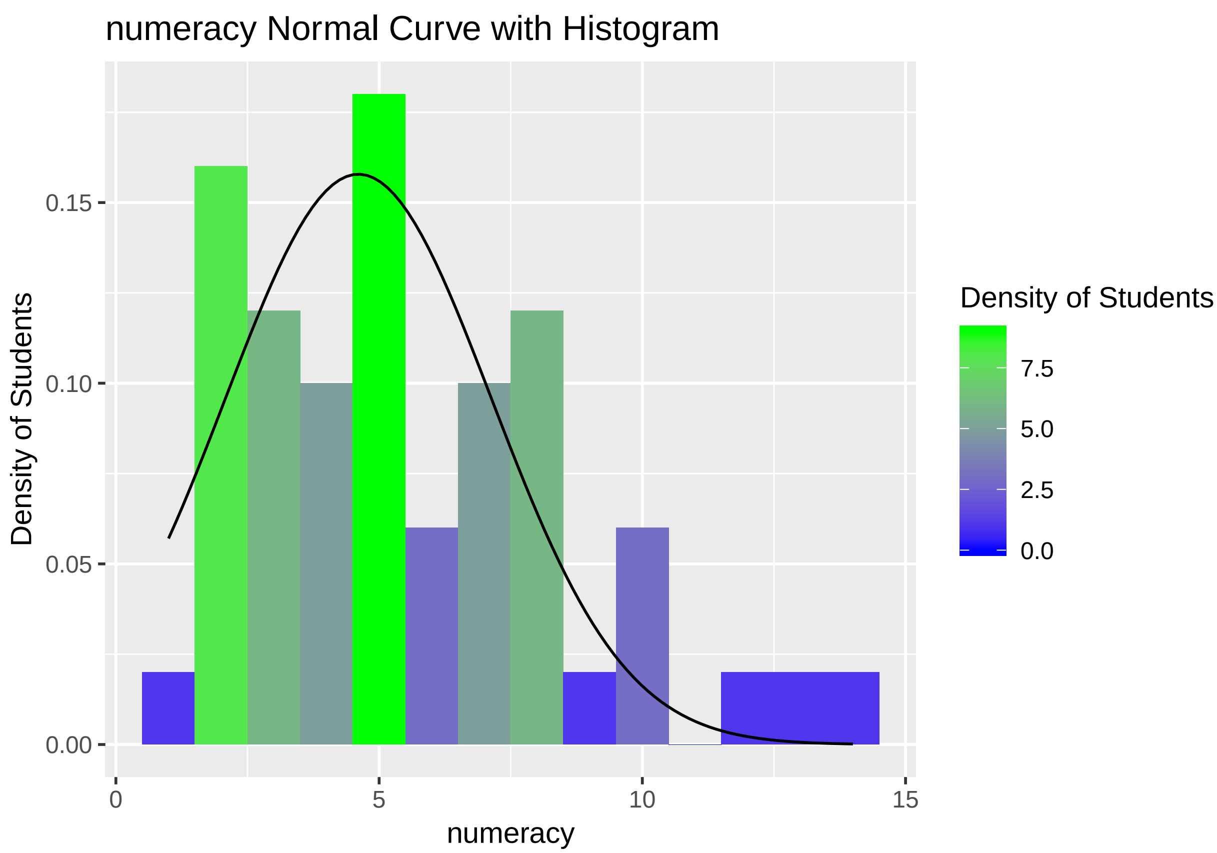

Lectures Scores
Number of Students
6
4
2
0
25
50
75
100
Percent Achieved
Number of Students
6
4
2
0


lectures Normal Curve with Histogram
Density of Students
0.02
0.01
0.00
25
50
75
100
lectures
Density of Students
6
4
2
0

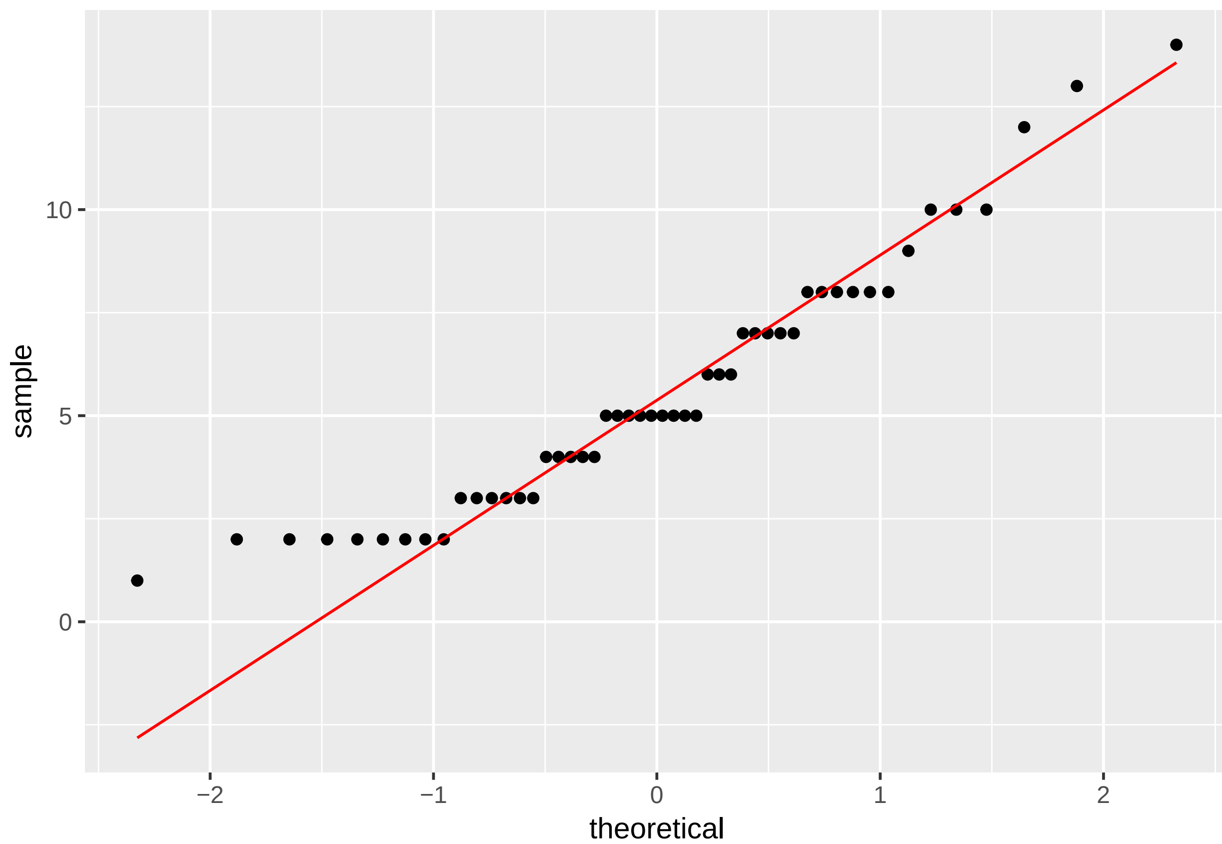


sample
90
60
30
0
−2
−1
0
1
2
theoretical

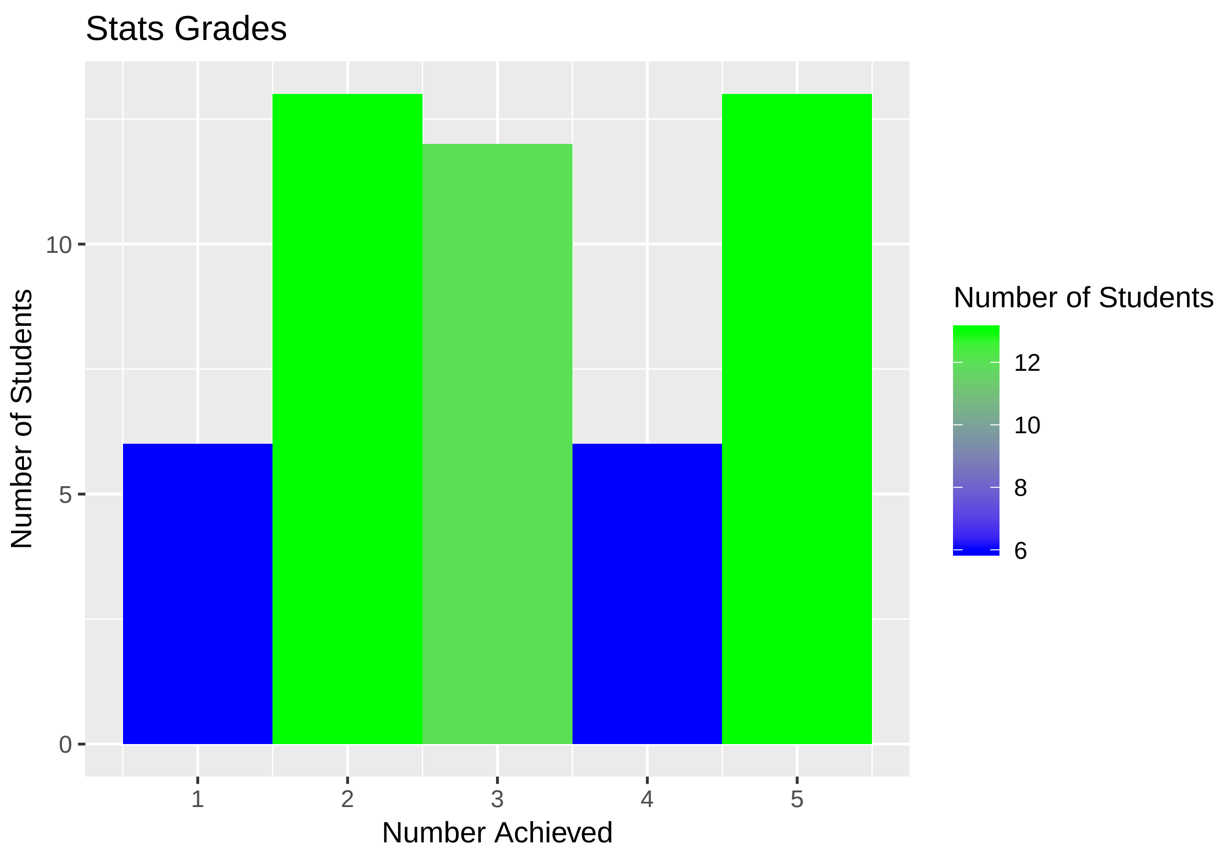

Numeracy Scores
Number of Students
Number Achieved
0.0
2.5
5.0
7.5
0
5
10
15
Number of Students
7.5
5.0
2.5
0.0

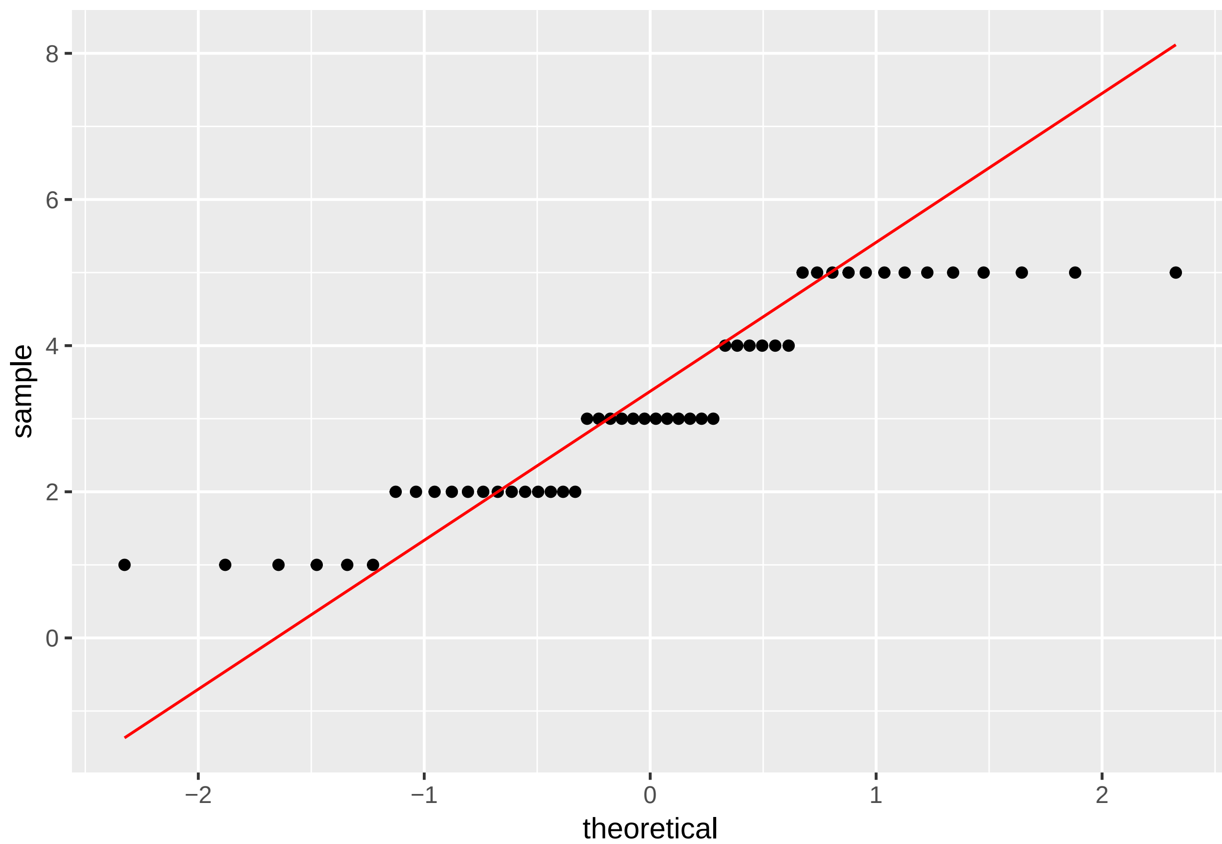


numeracy Normal Curve with Histogram
Density of Students
0.00
0.05
0.10
0.15
0
5
10
15
numeracy
Density of Students
0.0
2.5
5.0
7.5



10
5
0
sample
−2
−1
0
1
2
theoretical


Stats Grades
Number of Students
10
5
0
1
2
3
4
5
Number Achieved
Number of Students
12
10
8
6




```
#run statistics for each of the evaluations
examStatsJeff     <- stat.desc(examDataJeff$exam, norm = TRUE)
computerStatsJeff <- stat.desc(examDataJeff$computer, norm = TRUE)
lecturesStatsJeff <- stat.desc(examDataJeff$lectures, norm = TRUE)
numeracyStatsJeff <- stat.desc(examDataJeff$numeracy,norm = TRUE)
statsStatsJeff    <- stat.desc(examDataJeff$stats,norm = TRUE)

examStatsJeff
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##        50.00         0.00         0.00        56.00        99.00
##        range          sum       median         mean      SE.mean
##        43.00      3801.00        75.00        76.02         1.44
## CI.mean.0.95          var      std.dev     coef.var     skewness
##         2.90       104.14        10.21         0.13         0.26
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##         0.38        -0.46        -0.35         0.98         0.72
```

```
computerStatsJeff
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##       50.000        0.000        0.000       27.000       73.000
##        range          sum       median         mean      SE.mean
##       46.000     2558.000       54.000       51.160        1.203
## CI.mean.0.95          var      std.dev     coef.var     skewness
##        2.417       72.341        8.505        0.166       -0.506
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
```

```
##       -0.752         0.964         0.728         0.944         0.019
```

```r
lecturesStatsJeff
```

```
##       nbr.val     nbr.null       nbr.na          min          max
##         50.00         0.00         0.00        12.50       100.00
##         range          sum       median         mean      SE.mean
##         87.50      3163.50        65.75        63.27         2.68
## CI.mean.0.95          var      std.dev     coef.var     skewness
##          5.39       359.85        18.97         0.30        -0.34
##      skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##         -0.51        -0.42        -0.32         0.98         0.63
```

```r
numeracyStatsJeff
```

```
##       nbr.val     nbr.null       nbr.na          min          max
##       50.0000       0.0000       0.0000       1.0000      14.0000
##         range          sum       median         mean      SE.mean
##       13.0000     279.0000       5.0000       5.5800       0.4343
## CI.mean.0.95          var      std.dev     coef.var     skewness
##        0.8728       9.4322       3.0712       0.5504       0.7464
##      skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##        1.1087      -0.0064      -0.0049       0.9323       0.0068
```

```r
statsStatsJeff
```

```
##       nbr.val     nbr.null       nbr.na          min          max
##      50.00000      0.00000      0.00000      1.00000      5.00000
##         range          sum       median         mean      SE.mean
##       4.00000    157.00000      3.00000      3.14000      0.19590
## CI.mean.0.95          var      std.dev     coef.var     skewness
##       0.39367      1.91878      1.38520      0.44115      0.07054
##      skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##       0.10478     -1.33406     -1.00774      0.88013      0.00011
```

```r
shapiro.test(examDataJeff$exam)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  examDataJeff$exam
## W = 1, p-value = 0.7
```

```r
shapiro.test(examDataJeff$computer)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  examDataJeff$computer
## W = 0.9, p-value = 0.02
```

```r
shapiro.test(examDataJeff$lectures)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  examDataJeff$lectures
## W = 1, p-value = 0.6
```

```r
shapiro.test(examDataJeff$numeracy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataJeff$numeracy
## W = 0.9, p-value = 0.007
```

```r
shapiro.test(examDataJeff$stats)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataJeff$stats
## W = 0.9, p-value = 0.0001
```

What do the Professor Jeff results show?

First, a bit about the students themselves. 1. 50 of the students in the cohort of 150 were in class with Professor Jeff. 2. Unlike the Professor George cohort, all students in the Professor Jeff cohort have grades recorded (stats variable).

Similarly to the Professor George cohort, the numerical evaluation score histograms generally show that being in class with Professor Jeff has an effect on student results, although this cohort's results are much more favorable than the previous one analyzed.

1. The exam results are generally positive, with the mean in the high 70%'s, the range between 50-100%, and the data normally distributed.
2. The computer results are not as normally distributed, instead showing a significant left skew, with only a few results to the right of the largest data groupings, which sit between 50-60%. There are only a few results higher than 60%, and many below 50%.
3. The lectures score are evenly distributed, with the vast majority between around 40-80%, but without a normalized grouping around any range of values. Because of the way that the data is distributed, I set the binwidth to 5, having the histogram draw its groupings in 5% sizes. This makes the graph easier to read, confirming this conclusion.
4. The numeracy scores center around 5, but differently from the previous cohort, several of Professor Jeff's students did very well, scoring between 10 and 15. Again, though, the numeracy scores are generally skewed to the low end of the range, indicating that the students in general did not do well on this evaluation.

Additionally, looking at the final student grades within the Professor Jeff cohort, there are two conclusions from the histogram.

1. The grades distribution for this cohort looks somewhat similar to that of the overall student population, leading us to initially conclude that the final measure of student success is not positively or negatively affected by Professor Jeff's being the instructor.
2. A high number of student (12 of 50) failed the class.

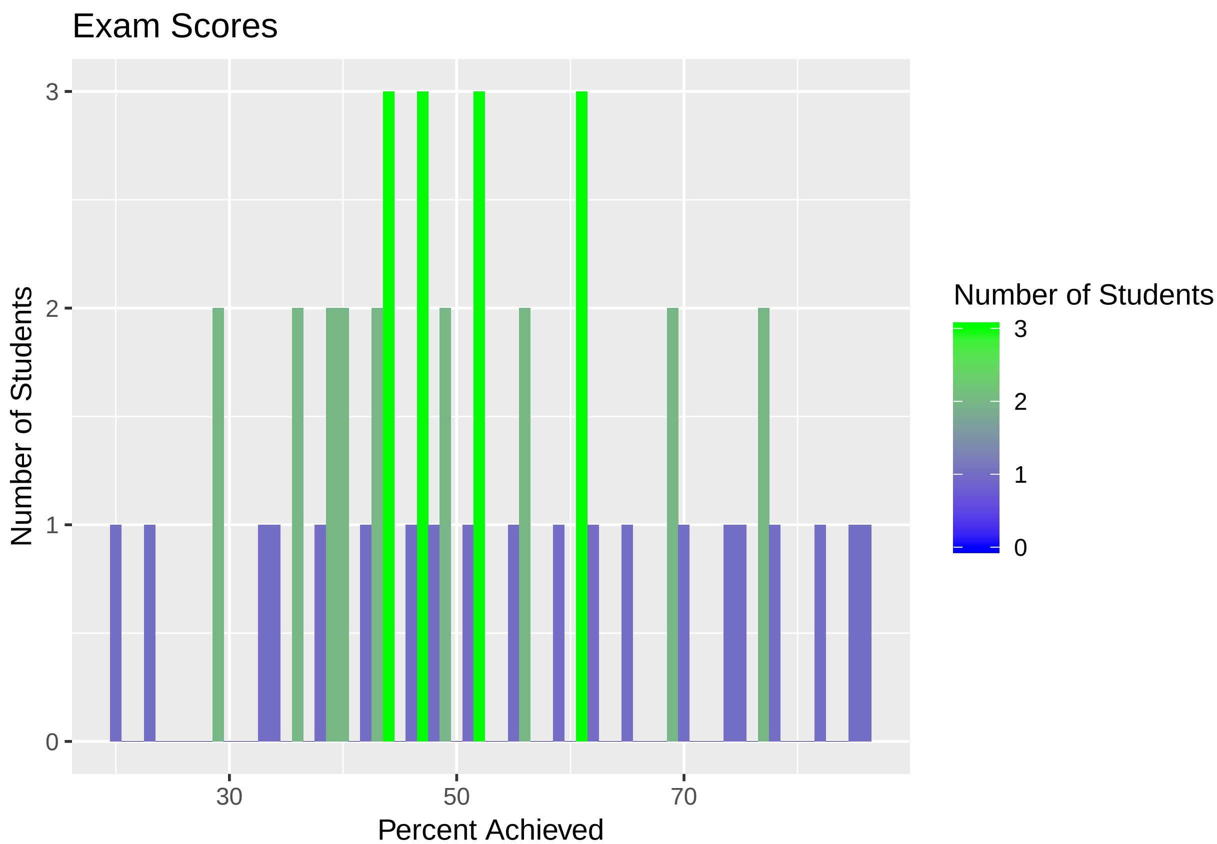
Finally, looking at the statistics confirms the conclusions that I reached from the visual representations.

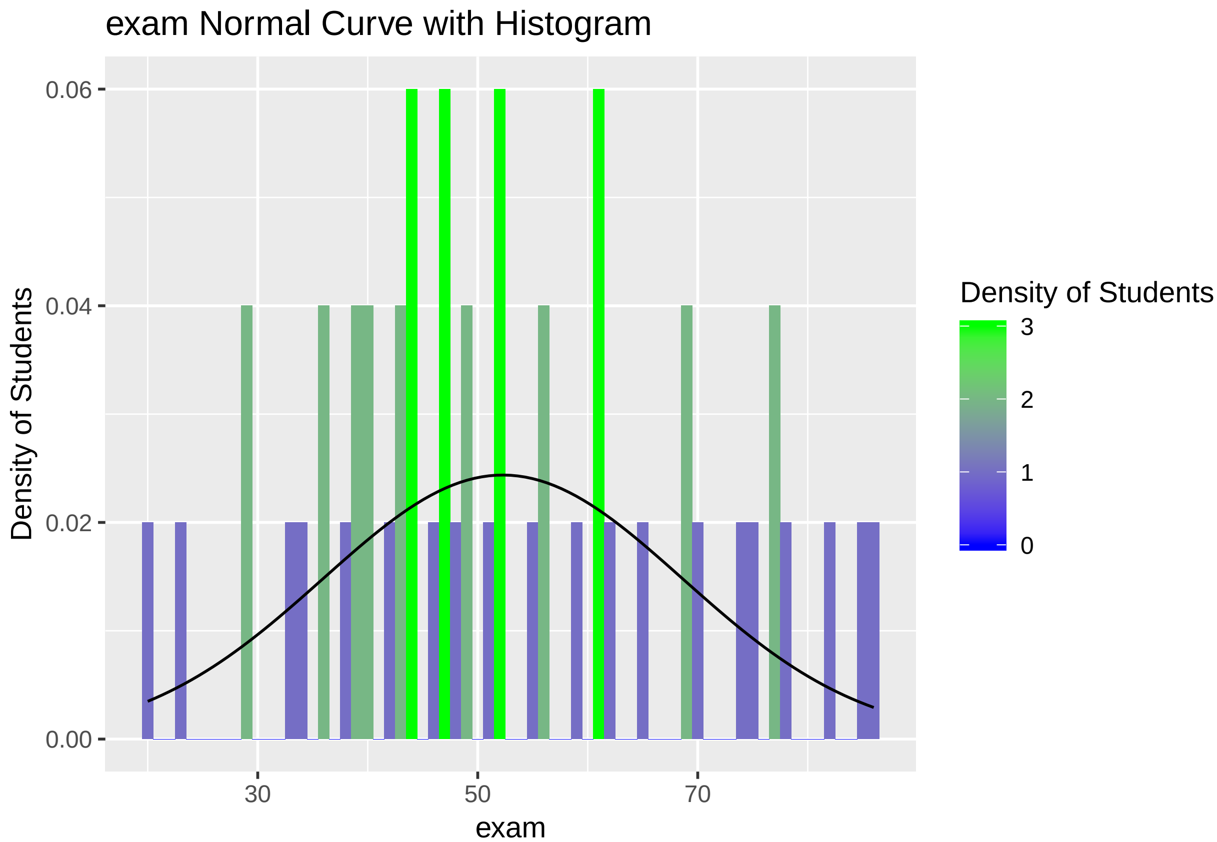
1. The means and medians confirm that the students performed better in this cohort than the previous one, with the exam and numeracy scores in particular being higher, while the standard deviations confirm the spread of the results.
2. The skew and kurtosis values again confirm the results seen in the histograms.
3. The Shapiro-Wilk tests for exam, and lectures all confirm that their scores are normally distributed, while the same tests for computer, numeracy, and stats confirm that those scores are not normally distributed. Again, this confirms what we are seeing graphically.

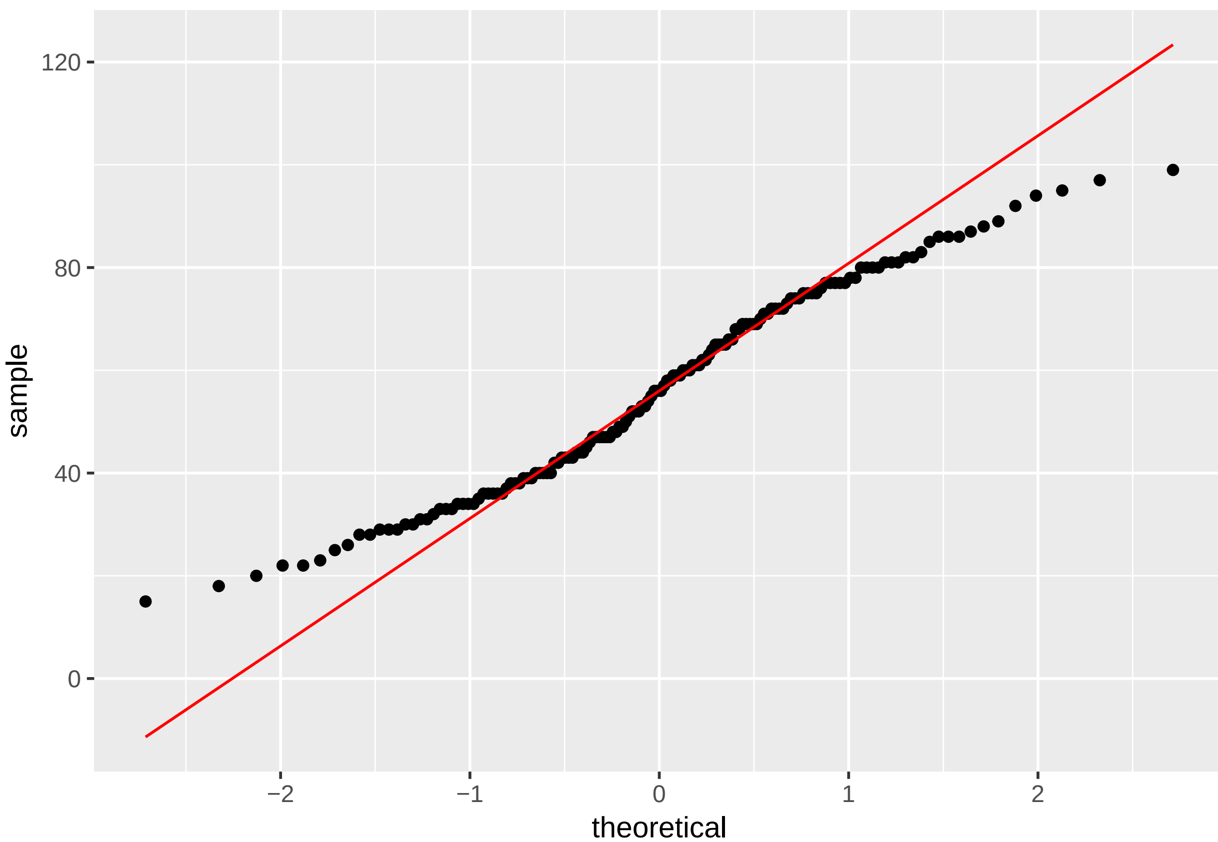
Finally, let's examine Professor Tushmann's results.

```
#subset the data for Professor Tushmann
examDataTushmann  <- subset(examData, uni=="2",
                            select=c("exam", "computer", "lectures", "numeracy", "stats"))
str(examDataTushmann)
```

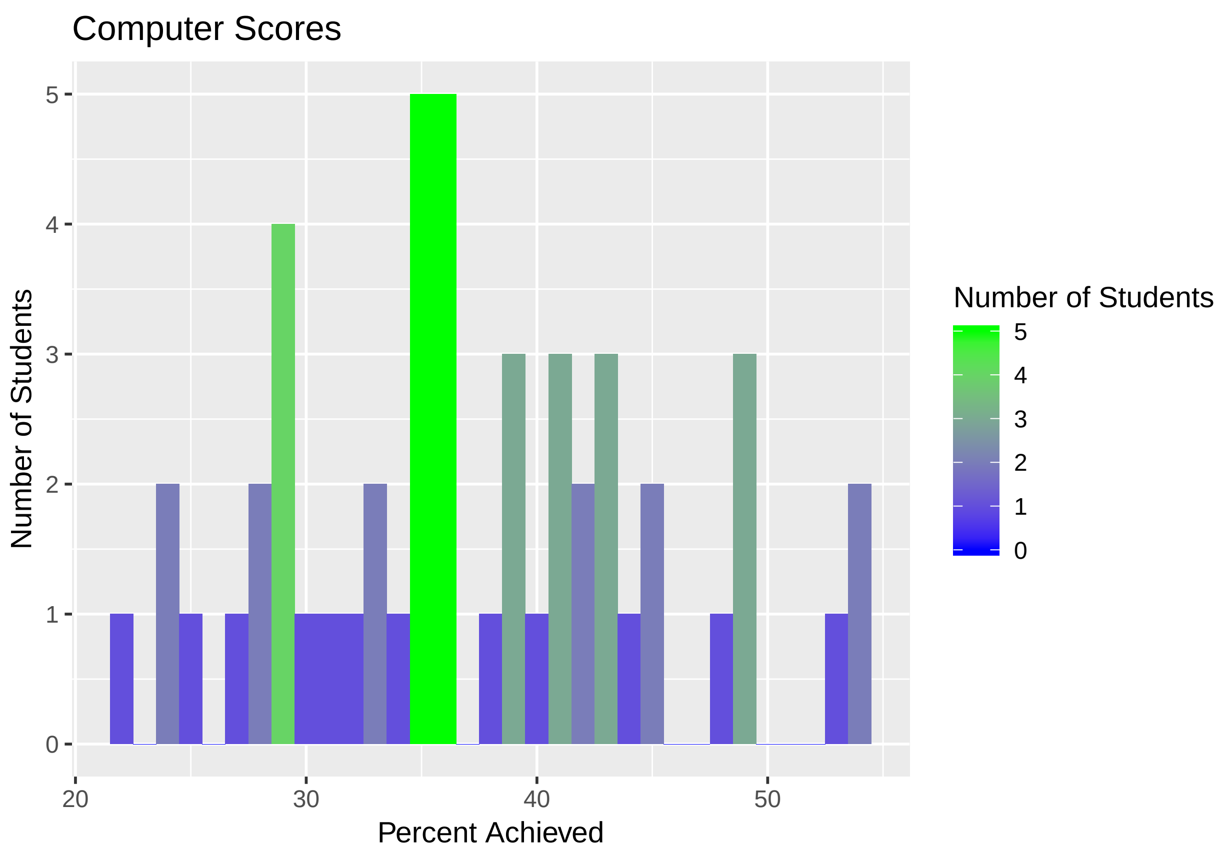
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    50 obs. of  5 variables:
##  $ exam    : num  23 39 52 39 52 20 47 52 82 40 ...
##  $ computer: num  41 34 45 24 40 35 36 36 32 49 ...
##  $ lectures: num  75 8.5 69.5 67 44.5 76.5 70 18.5 43.5 100 ...
##  $ numeracy: num  7 1 6 6 2 8 3 7 4 6 ...
##  $ stats   : num  1 1 1 2 2 2 2 3 3 3 ...
```

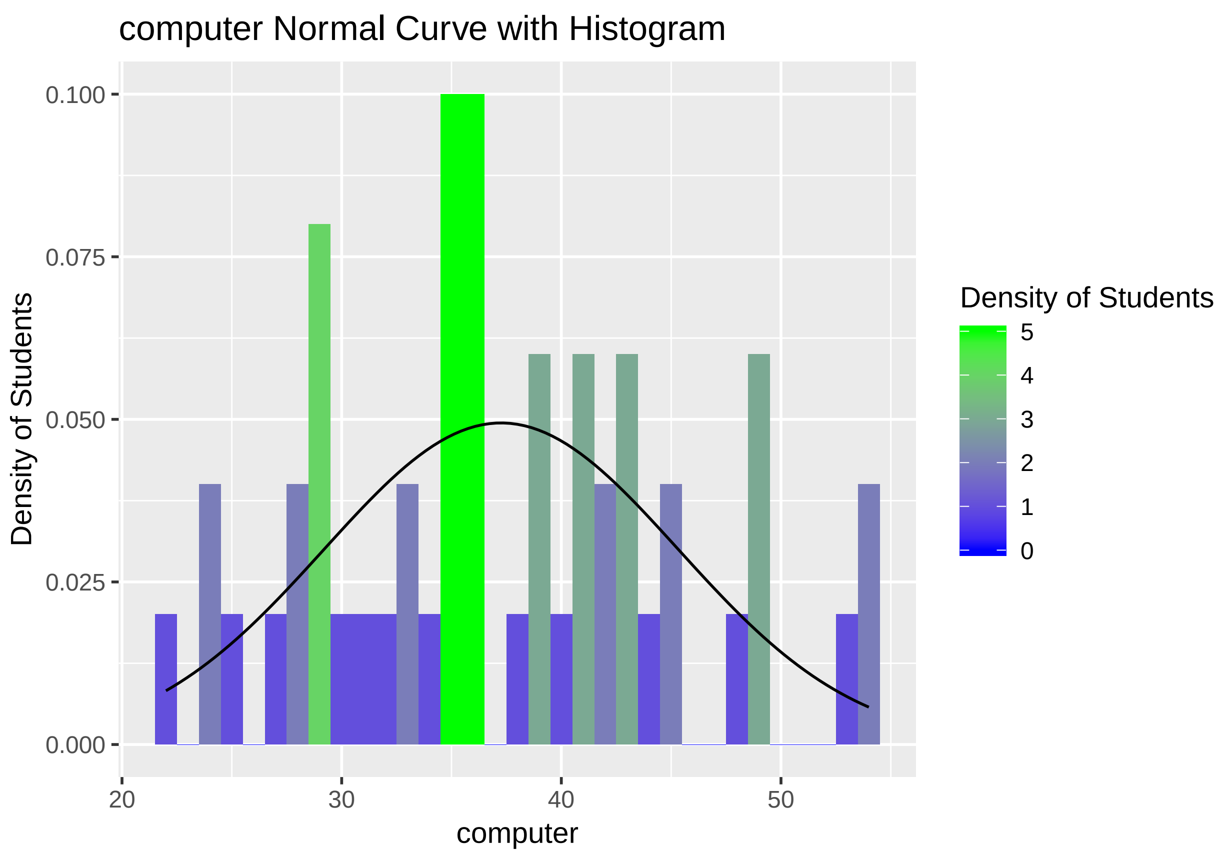


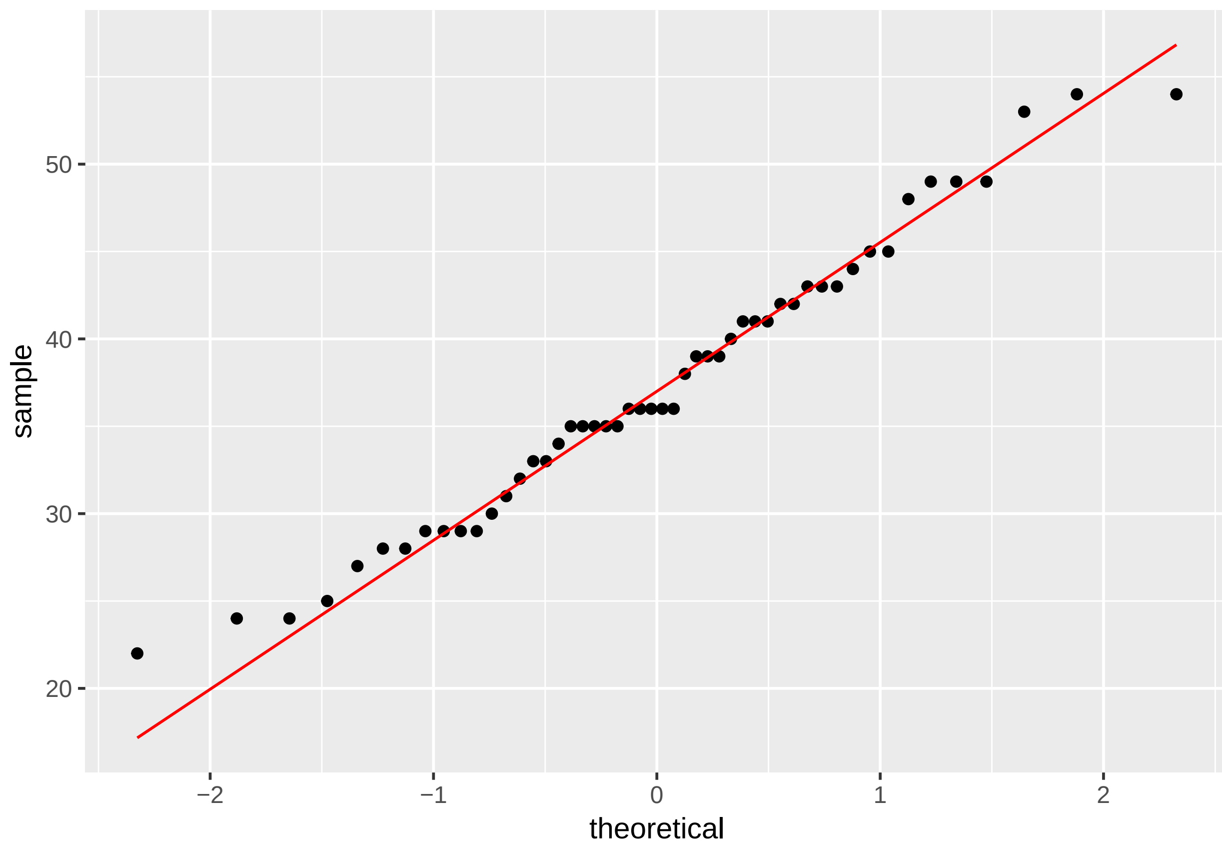

exam Normal Curve with Histogram
Density of Students
0.06
0.04
0.02
0.00
30
50
70
exam
Density of Students
3
2
1
0

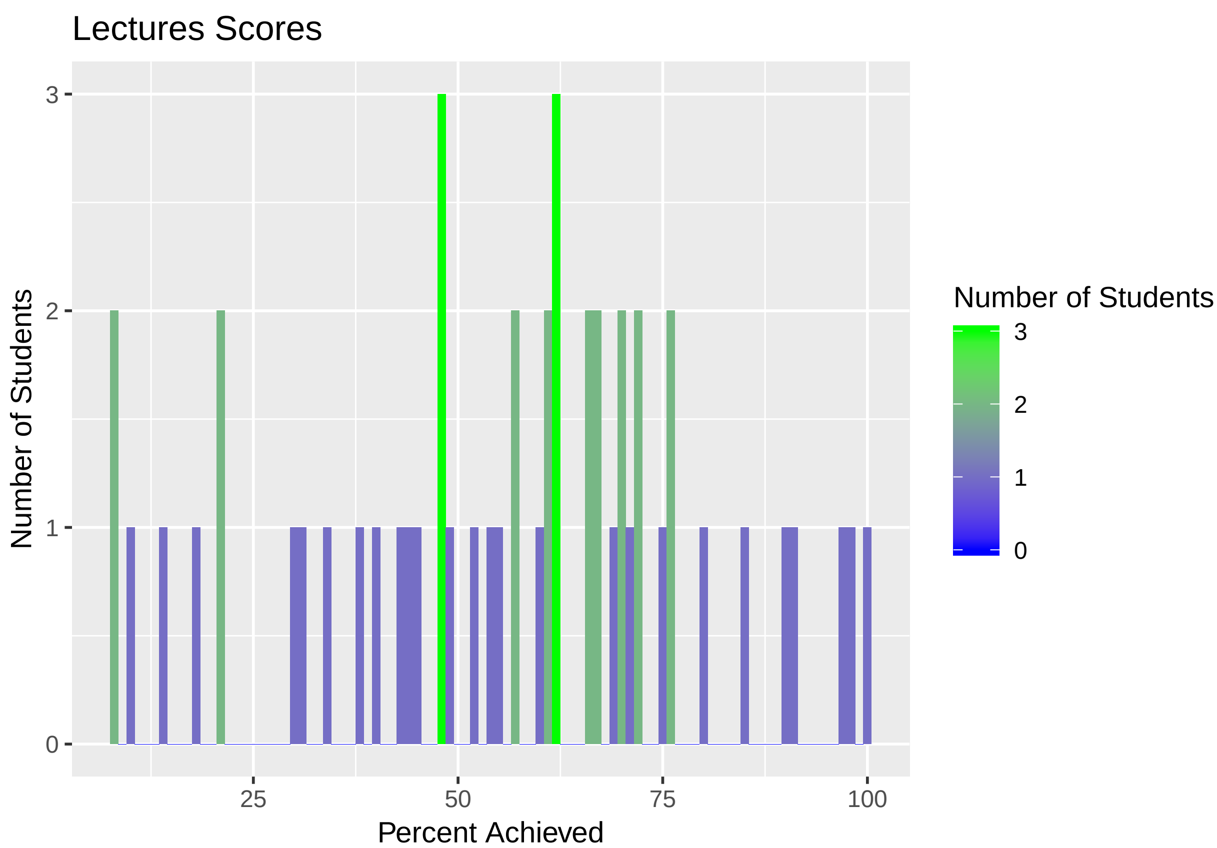


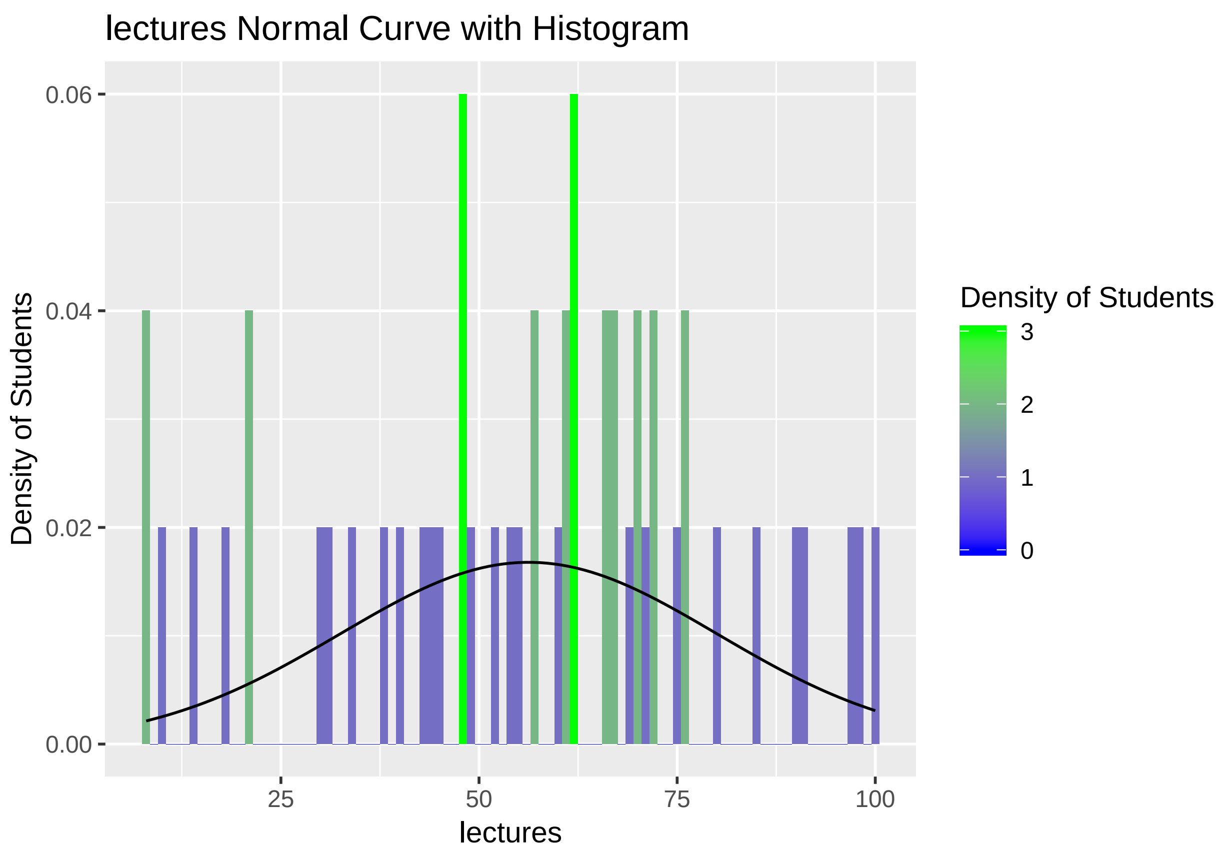
120
80
40
0
−2
−1
0
1
2
theoretical
sample

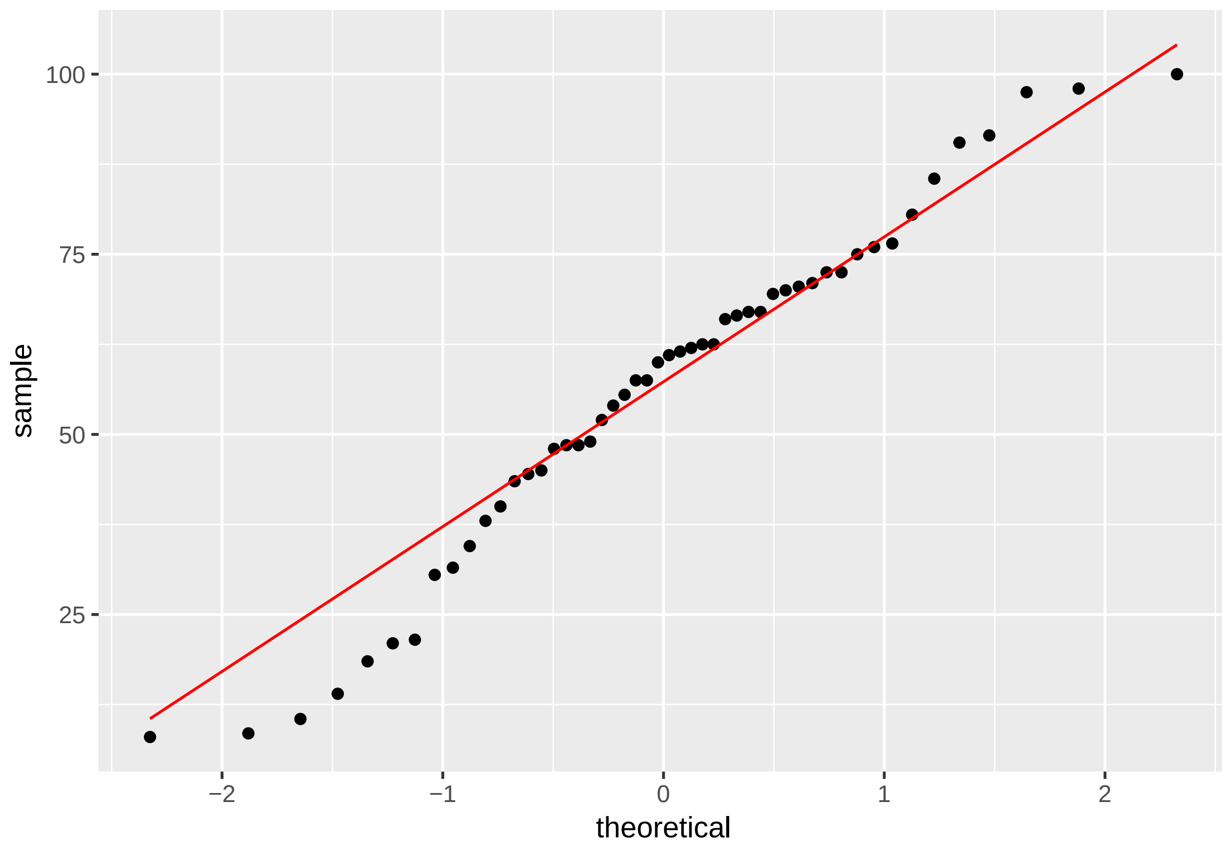

Computer Scores
Number of Students
0
1
2
3
4
5
20
30
40
50
Percent Achieved
Number of Students
5
4
3
2
1
0

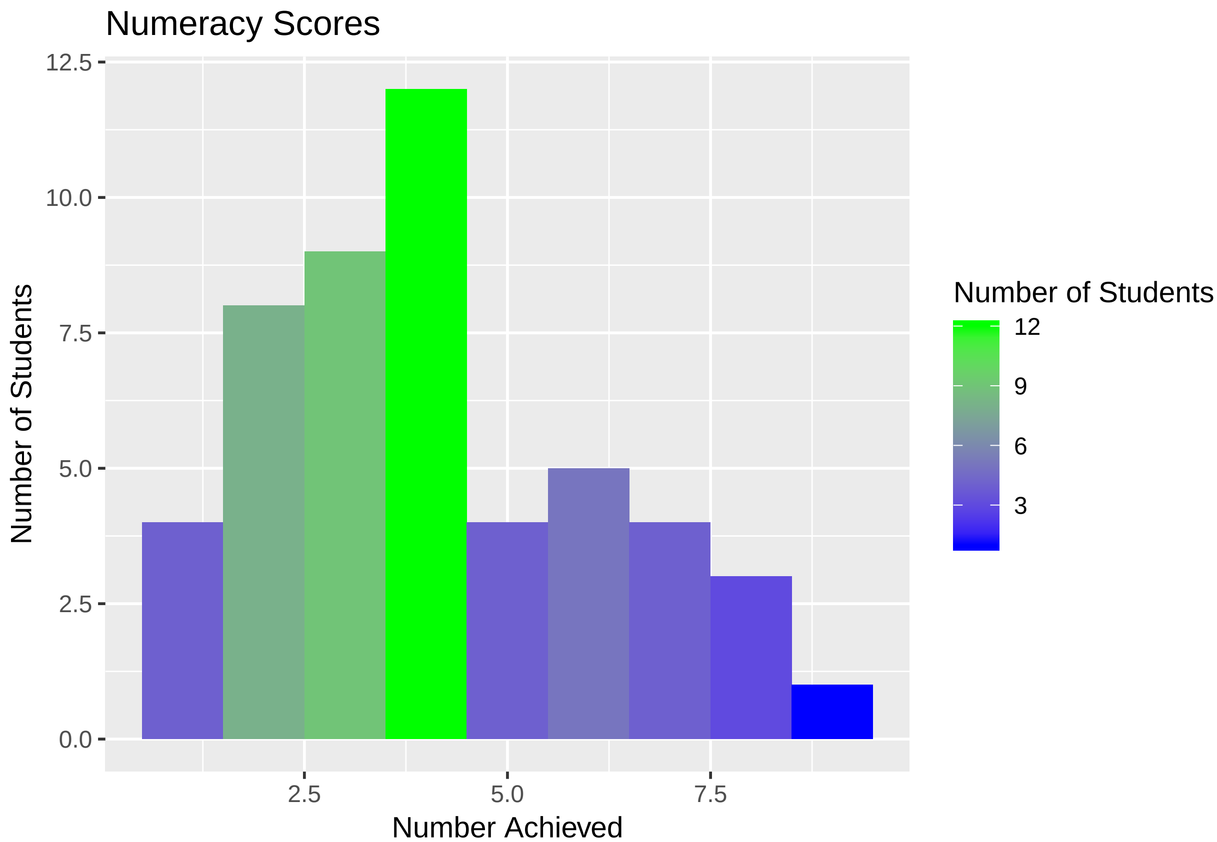

computer Normal Curve with Histogram
Density of Students
0.100
0.075
0.050
0.025
0.000
20
30
40
50
computer
Density of Students
5
4
3
2
1
0

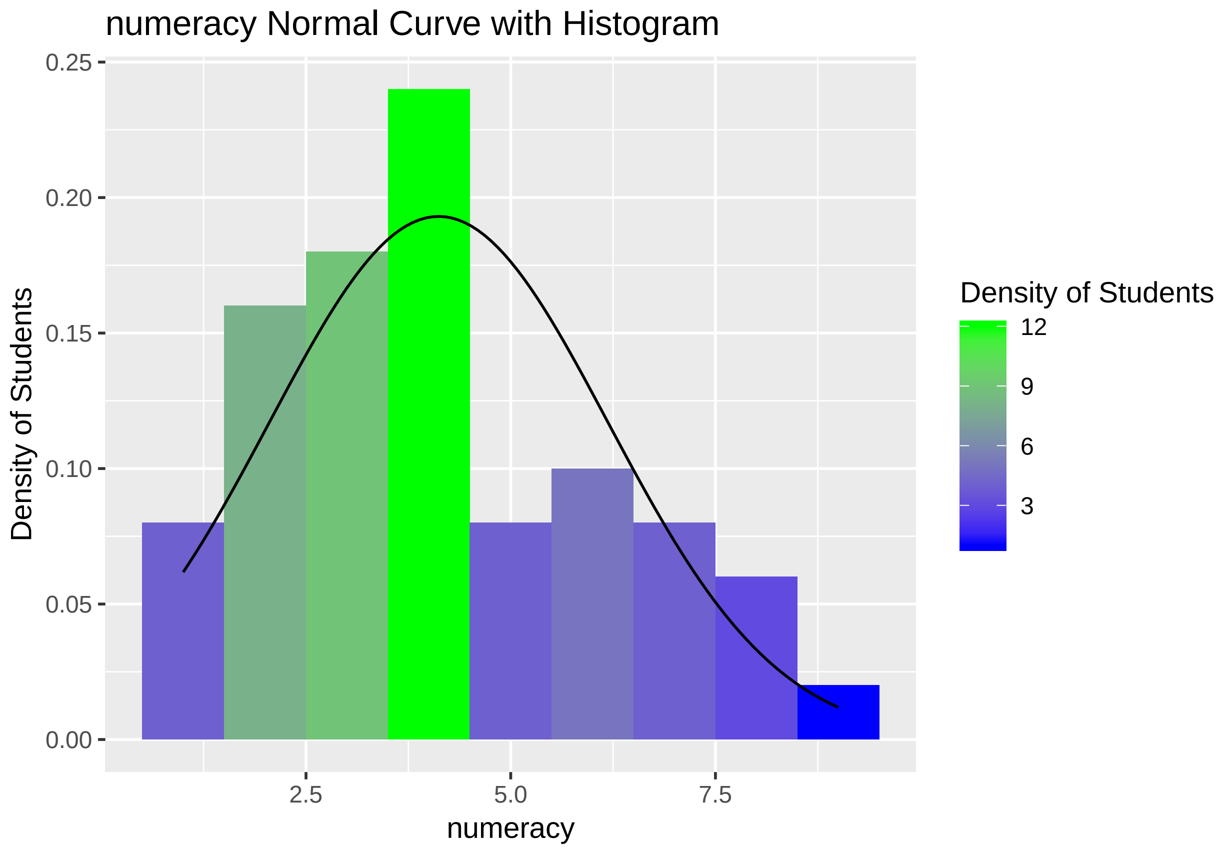

50
40
30
20
sample
−2
−1
0
1
2
theoretical

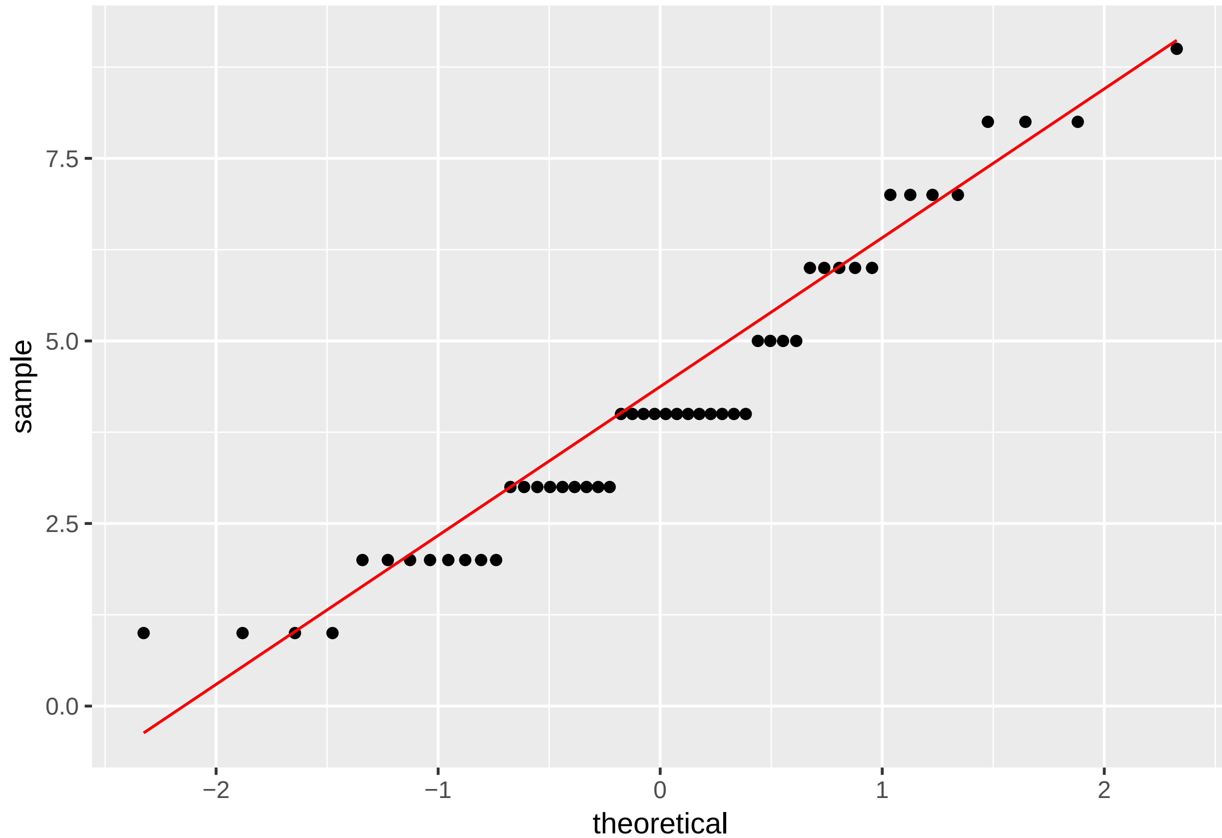

Lectures Scores
Number of Students
Percent Achieved
0
1
2
3
25
50
75
100
Number of Students
3
2
1
0


lectures Normal Curve with Histogram
Density of Students
0.06
0.04
0.02
0.00
25
50
75
100
lectures
Density of Students
3
2
1
0

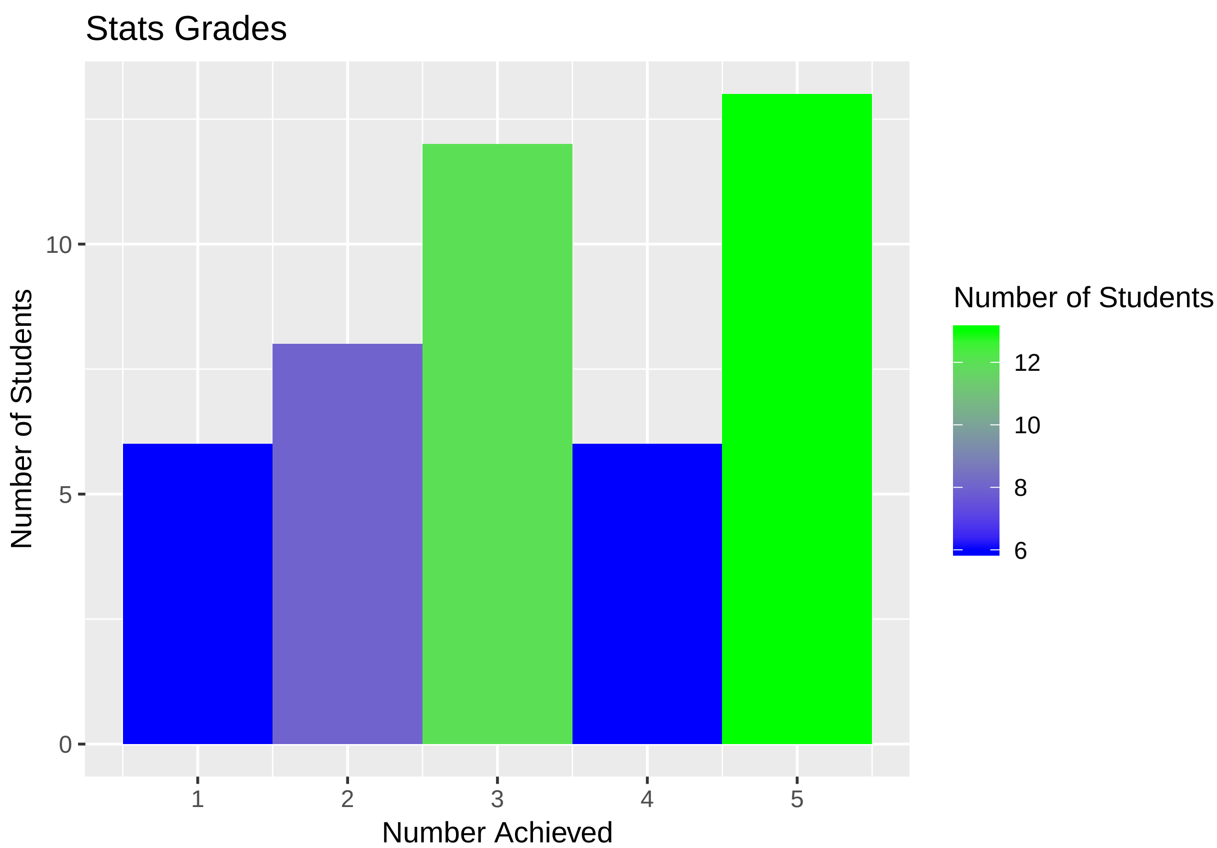

100
75
50
25
sample
−2
−1
0
1
2
theoretical


Numeracy Scores
12.5
10.0
7.5
5.0
2.5
0.0
Number of Students
2.5
5.0
7.5
Number Achieved
Number of Students
12
9
6
3

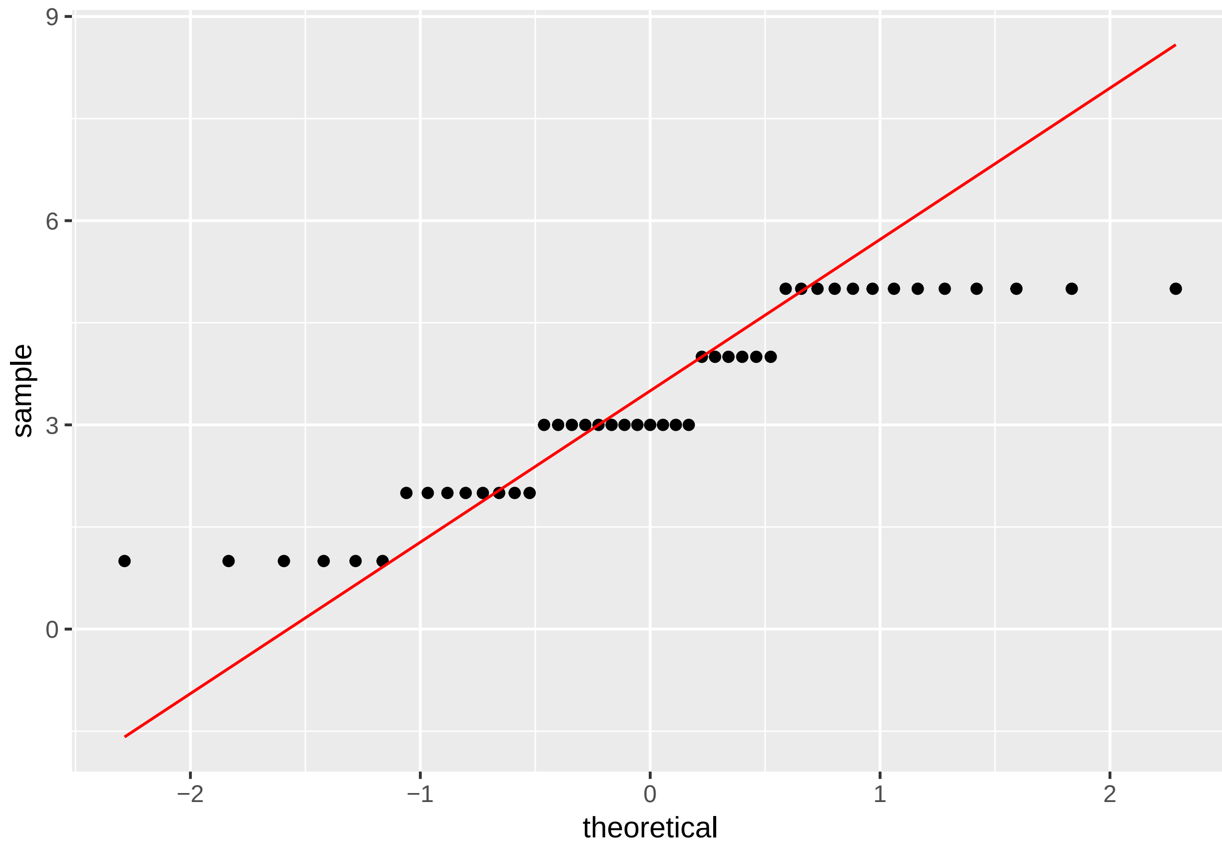

numeracy Normal Curve with Histogram
0.25
0.20
0.15
0.10
0.05
0.00
Density of Students
2.5
5.0
7.5
numeracy
Density of Students
12
9
6
3




```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```




```
## Warning: Removed 5 rows containing non-finite values (stat_qq).
## Warning: Removed 5 rows containing non-finite values (stat_qq_line).
```




```r
#run statistics for each of the evaluations
examStatsTushmann     <- stat.desc(examDataTushmann$exam, norm = TRUE)
computerStatsTushmann <- stat.desc(examDataTushmann$computer, norm = TRUE)
lecturesStatsTushmann <- stat.desc(examDataTushmann$lectures, norm = TRUE)
numeracyStatsTushmann <- stat.desc(examDataTushmann$numeracy,norm = TRUE)
statStatsTushmann     <- stat.desc(examDataTushmann$stats,norm = TRUE)

examStatsTushmann
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##         50.00          0.00          0.00         20.00         86.00
##         range           sum        median          mean       SE.mean
##         66.00       2614.00         49.00         52.28          2.31
## CI.mean.0.95           var       std.dev      coef.var      skewness
##          4.65        267.88         16.37          0.31          0.30
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##          0.45         -0.72         -0.54          0.97          0.27
```

```r
computerStatsTushmann
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##         50.00          0.00          0.00         22.00         54.00
##         range           sum        median          mean       SE.mean
##         32.00       1863.00         36.00         37.26          1.14
## CI.mean.0.95           var       std.dev      coef.var      skewness
##          2.29         65.09          8.07          0.22          0.21
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
```

```
##         0.32        -0.68        -0.51         0.98         0.46
```

```r
lecturesStatsTushmann
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##        50.00         0.00         0.00         8.00       100.00
##        range          sum       median         mean      SE.mean
##        92.00      2813.00        60.50        56.26         3.36
## CI.mean.0.95          var      std.dev     coef.var     skewness
##         6.76       565.14        23.77         0.42        -0.29
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##        -0.43        -0.56        -0.43         0.97         0.23
```

```r
numeracyStatsTushmann
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##       50.000        0.000        0.000        1.000        9.000
##        range          sum       median         mean      SE.mean
##        8.000      206.000        4.000        4.120        0.292
## CI.mean.0.95          var      std.dev     coef.var     skewness
##        0.587        4.271        2.067        0.502        0.482
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##        0.715       -0.652       -0.492        0.941        0.015
```

```r
statStatsTushmann
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##     45.00000      0.00000      5.00000      1.00000      5.00000
##        range          sum       median         mean      SE.mean
##      4.00000    147.00000      3.00000      3.26667      0.20938
## CI.mean.0.95          var      std.dev     coef.var     skewness
##      0.42197      1.97273      1.40454      0.42996     -0.13068
##     skew.2SE     kurtosis     kurt.2SE   normtest.W   normtest.p
##     -0.18471     -1.29768     -0.93420      0.87938      0.00023
```

```r
shapiro.test(examDataTushmann$exam)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  examDataTushmann$exam
## W = 1, p-value = 0.3
```

```r
shapiro.test(examDataTushmann$computer)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  examDataTushmann$computer
## W = 1, p-value = 0.5
```

```r
shapiro.test(examDataTushmann$lectures)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  examDataTushmann$lectures
## W = 1, p-value = 0.2
```

```r
shapiro.test(examDataTushmann$numeracy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataTushmann$numeracy
## W = 0.9, p-value = 0.01
```

```r
shapiro.test(examDataTushmann$stats)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  examDataTushmann$stats
## W = 0.9, p-value = 0.0002
```

What do the Professor Tushmann results show?

First, a bit about the students themselves. 1. 50 of the students in the cohort of 150 were in class with Professor Tushmann. 2. Of those 50, 5 students do not have grades recorded (stats variable).

Similarly to the Professor George cohort, the numerical evaluation score histograms generally show that being in class with Professor Jeff has an effect on student results, although this cohort's results are much more favorable than the previous one analyzed.

1. The exam results are mixed, in that they appear to center around 45-55%, but their spread is large across the results, with results ranging from 20% to nearly 90%. The results appear to skew a bit on the lower end, with more below than above 50%.
2. The computer scores show a bit of an opposite distribution from the exam results, in that the results are skewed right versus left. The spread only goes from 20-60%, and the scores center around 32%. Overall, not good results for this evaluation.
3. The lectures scores are evenly distributed across the entire spectrum of possible results, from 0-100%. While the number of scores between 60% and 70& is a bit higher than the remainder of the range, it doesn't appear to be significant.
4. Numeracy scores for the cohort are not exceptionally good, with values at 4 and below comprising half of the results, and high score only 9. The distribution has a longer right tail, simply because there are more possible values between 4 and 9 than there are between 1 and 4.

Additionally, looking at the final student grades within the Professor Tushmann cohort, there are two conclusions from the histogram.

1. The grades distribution for this cohort looks somewhat similar to that of the overall student population, leading us to initially conclude that the final measure of student success is not positively or negatively affected by Professor Tushmann's being the instructor.
2. A high number of student (13 of 45 with grades recorded) failed the class.

Finally, looking at the statistics confirms the conclusions that I reached from the visual representations.

1. The means and medians confirm that the students performed better in this cohort than the previous one, with the exam and numeracy scores in particular being higher, while the standard deviations confirm the spread of the results.
2. The skew and kurtosis values again confirm the results seen in the histograms.

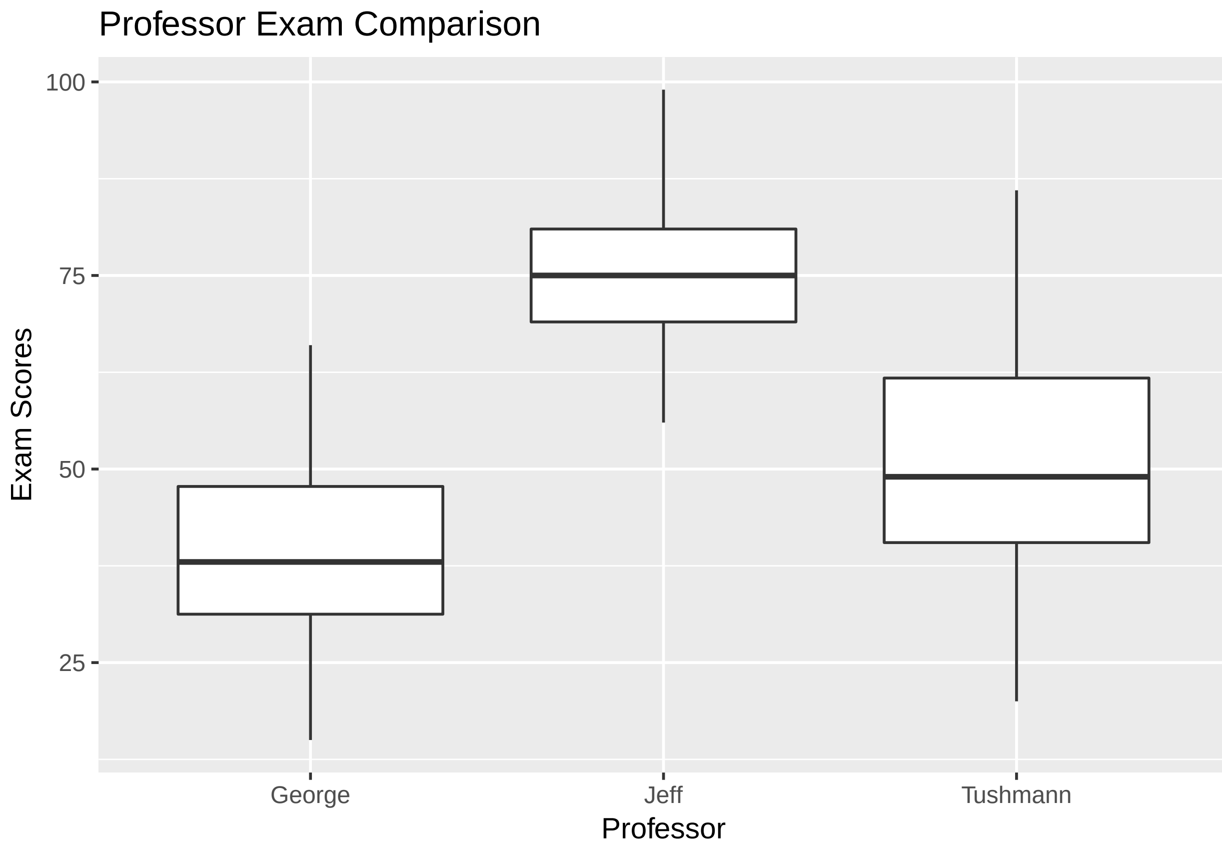
3. The Shapiro-Wilk tests for exam, computer, and lectures all confirm that their scores are normally distributed, while the same tests for numeracy and stats confirm that those scores are not normally distributed. Again, this confirms what we are seeing graphically.

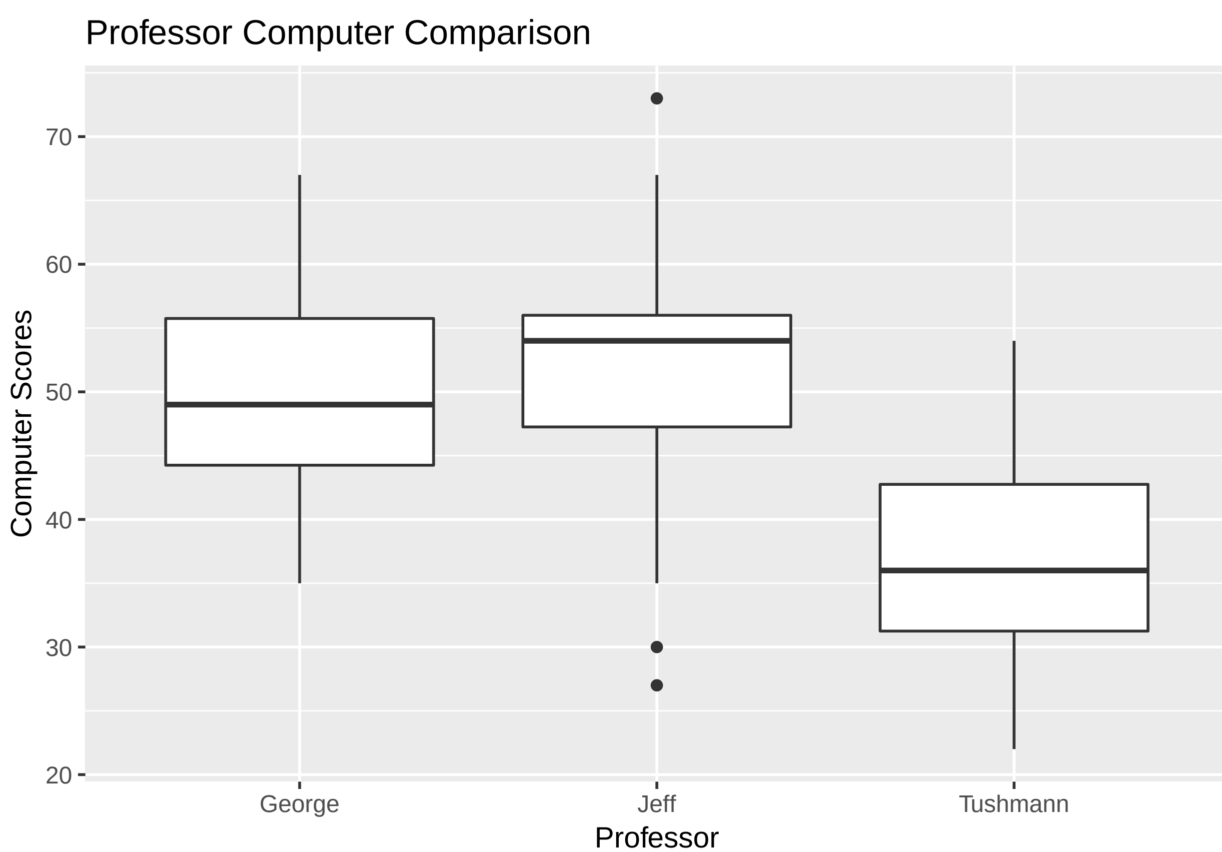
The Next Step: Professor Comparisons

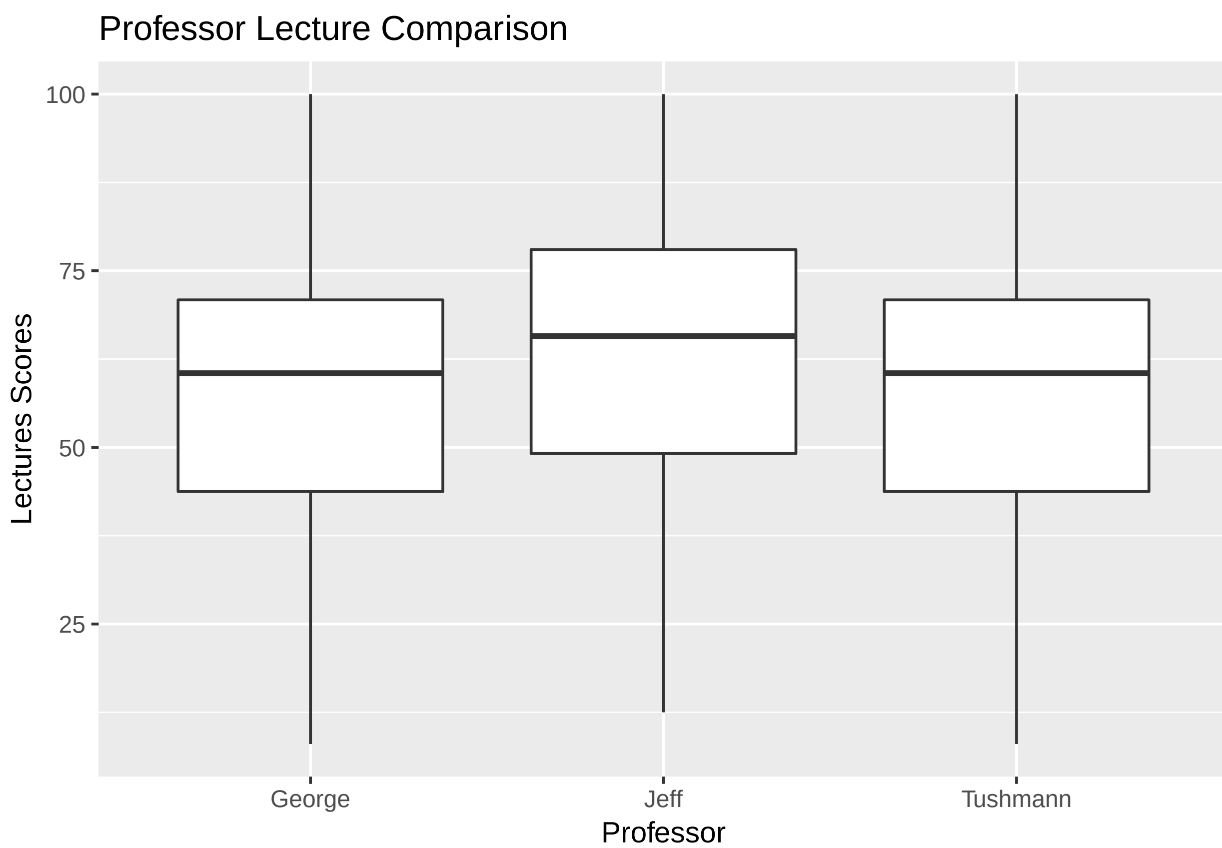
While there is interest in seeing how the students in each professor's section have performed, another, it is natural for me to want to compare how the professors have influenced the results of the students in their cohorts. In order to do so, we must make one assumption, that being that each of the student cohorts are representative of the student population at large, and there is not a bias in any of the cohorts. Practically, that would mean that the students were randomly assigned to their cohort, and that they had an equal chance of being assigned to each professor. While we don't absolutely know that this was the case, I will stipulate this for purposes of the analysis.

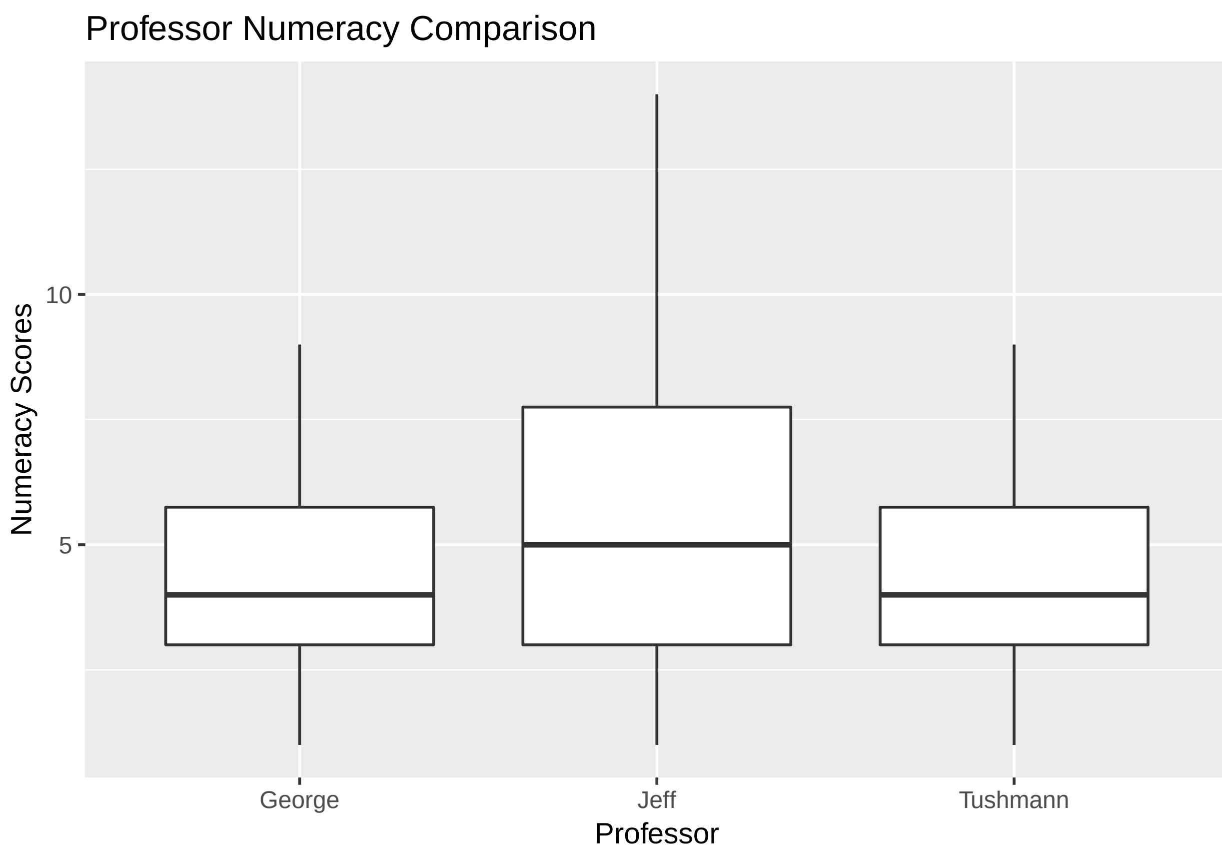
I will look at each of the 5 variables to determine if there are student performance differences between the professors, using box plots to visualize the data, and summary statistics to provide numerical verification of what we are seeing in the graphs.

First, the graphs

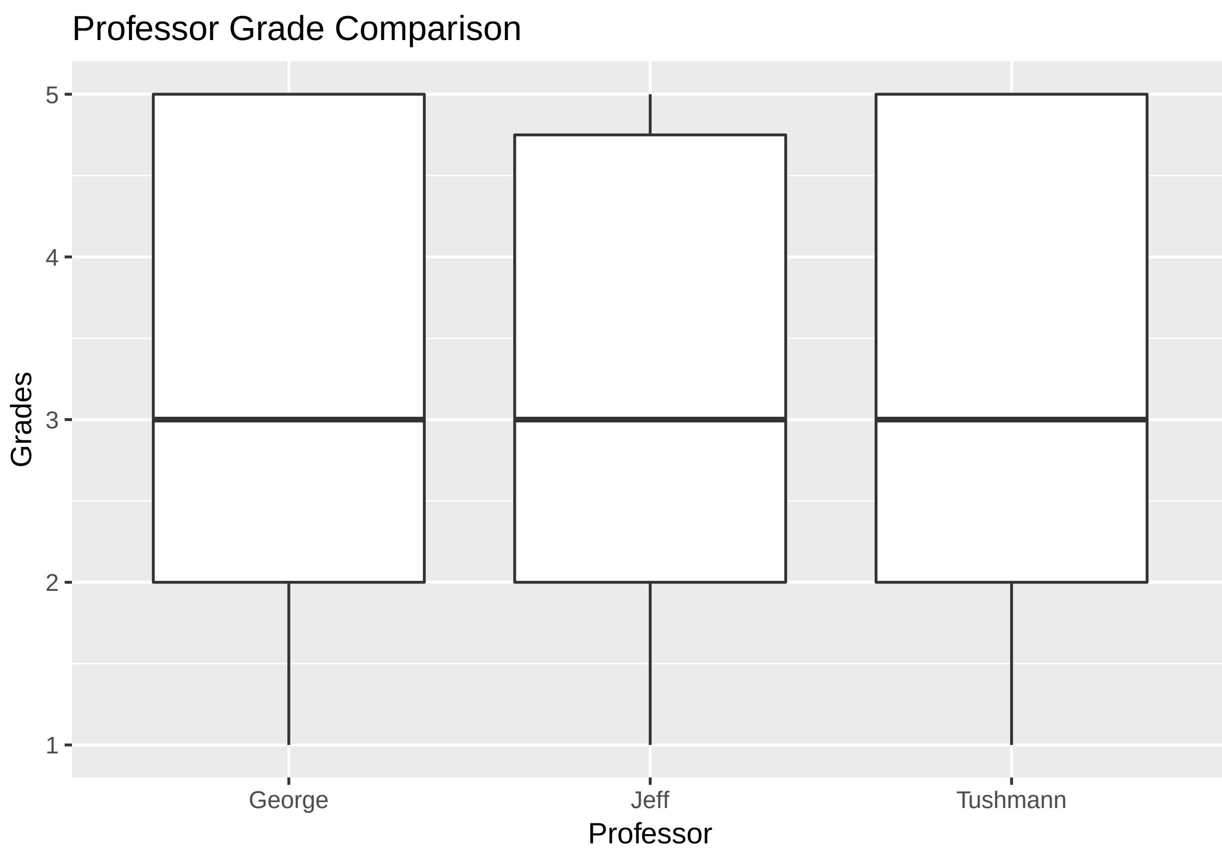





Professor Lecture Comparison
100
75
50
25
Lectures Scores
George
Jeff
Tushmann
Professor




```
## Warning: Removed 10 rows containing non-finite values (stat_boxplot).
```



Now, the summary statistics

```
describeBy(examData$exam, list(profUni),mat=TRUE,digits=3)
```

```
##     item   group1 vars  n mean sd median trimmed  mad min max range skew
## X11    1   George    1 50   40 13     38      40 12.6  15  66    51 0.29
## X12    2     Jeff    1 50   76 10     75      76  8.9  56  99    43 0.26
## X13    3 Tushmann    1 50   52 16     49      52 15.6  20  86    66 0.30
##     kurtosis  se
## X11    -0.72 1.8
## X12    -0.46 1.4
## X13    -0.72 2.3
```

```
describeBy(examData$computer, list(profUni),mat=TRUE,digits=3)
```

```
##     item   group1 vars  n mean  sd median trimmed mad min max range  skew
## X11    1   George    1 50   50 8.1     49      50 8.9  35  67    32  0.21
## X12    2     Jeff    1 50   51 8.5     54      52 5.9  27  73    46 -0.51
## X13    3 Tushmann    1 50   37 8.1     36      37 8.9  22  54    32  0.21
##     kurtosis  se
## X11    -0.68 1.1
## X12     0.96 1.2
## X13    -0.68 1.1
```

```
describeBy(examData$lectures, list(profUni),mat=TRUE,digits=3)
```

```
##     item   group1 vars  n mean sd median trimmed mad min max range  skew
## X11    1   George    1 50   56 24     60      57  20   8 100    92 -0.29
## X12    2     Jeff    1 50   63 19     66      64  21  12 100    88 -0.34
## X13    3 Tushmann    1 50   56 24     60      57  20   8 100    92 -0.29
##     kurtosis  se
## X11    -0.56 3.4
## X12    -0.42 2.7
## X13    -0.56 3.4
```

```
describeBy(examData$numeracy, list(profUni),mat=TRUE,digits=3)
```

```
##     item   group1 vars  n mean  sd median trimmed mad min max range skew
## X11    1   George    1 50  4.1 2.1      4     4.0 2.2   1   9     8 0.48
## X12    2     Jeff    1 50  5.6 3.1      5     5.3 3.0   1  14    13 0.75
## X13    3 Tushmann    1 50  4.1 2.1      4     4.0 2.2   1   9     8 0.48
##     kurtosis   se
## X11   -0.652 0.29
## X12   -0.006 0.43
## X13   -0.652 0.29
```

```
describeBy(examData$stats, list(profUni),mat=TRUE,digits=3)
```

```
##     item   group1 vars  n mean  sd median trimmed mad min max range   skew
## X11    1   George    1 45  3.3 1.4      3     3.3 1.5   1   5     4 -0.131
## X12    2     Jeff    1 50  3.1 1.4      3     3.2 1.5   1   5     4  0.071
## X13    3 Tushmann    1 45  3.3 1.4      3     3.3 1.5   1   5     4 -0.131
##     kurtosis   se
## X11     -1.3 0.21
## X12     -1.3 0.20
## X13     -1.3 0.21
```

So how do the professors compare to each other, in terms of student performance?

Exam

The exam scores are significantly different between the professors. Jeff's students did by far the best, followed by Tushmann's, and then George's. The box plot graph clearly shows this visually, and the summary statistics reinforce this conclusion. Jeff's students have by far the highest mean score, at 76%, while Tushmann's and George's are 52% and 40%, respectively. Additionally, the minimum score in the Jeff cohort is 56%, which is higher than the MEAN score in either of the other two cohorts.

Computer

The computer scores for George and Jeff are similar, while the Tushmann scores lag behind the other two. The distributions of scores between George and Jeff are bit different, with Jeff's middle 50% of scores more tightly grouped in the 3rd quartile, while George's are more evenly distributed. But their ranges are similar (excluding Jeff's outliers at each end of the range), and their means are within 1.5% of each other. The Jeff scores have a left skew, as we can see both in the box plot and with the skew value (-0.506), while the George ones have a slight right skew (0.212). Given these results, Jeff comes out slightly ahead in student performance, but not by much, compared to George.

Lectures and Numeracy

The lectures and numeracy scores show only slight differences between the professors, with Jeff's scores a bit better than the others, and the George and Tushmann scores identically distributed. Given this result, were I to have access to the actual professors and students, I would want to take a deeper look at how these results were generated, but given what we have, we have to accept the results as is.

Stats (Grades)

None of the professors excels in the distributions of their final student grades. Again, the George and Tushmann grade distributions are identical, and the Jeff distribution shows only a marginal improvement. While the numerical representations of the grades are integers, it is appropriate in this case to treat the variables as continuous, versus categorical, so that we can compare the results between the cohorts. As the box plots show, Jeff had slightly fewer students fail, otherwise it is difficult to separate cohort performance. The summary statistics reinforce this conclusion, as the mean, skew, and kurtosis results (of the cohorts) are all within 0.15 of each other. If the final grade is the determining factor in measuring the student success by who taught, the professors do not separate from each other, and I can conclude that the choice of professor is not material to the student's final grade.

Final Note

One final note that I would like call out is that the assignment called for me to do any data transformations, as I felt appropriate. As the above analysis shows, I did not perform any transformations on the data, because I did not feel that doing so would provide me with any information that would significantly enhance my analysis and understanding of student performance, where the variables did not conform to a normal distribution.

The only real data anomalies that I found were in the cases where the George and Tushmann results yielded identical summary statistics, and I feel that attempting to normalize those scores would not gain insights that would lead to any useful conclusions.