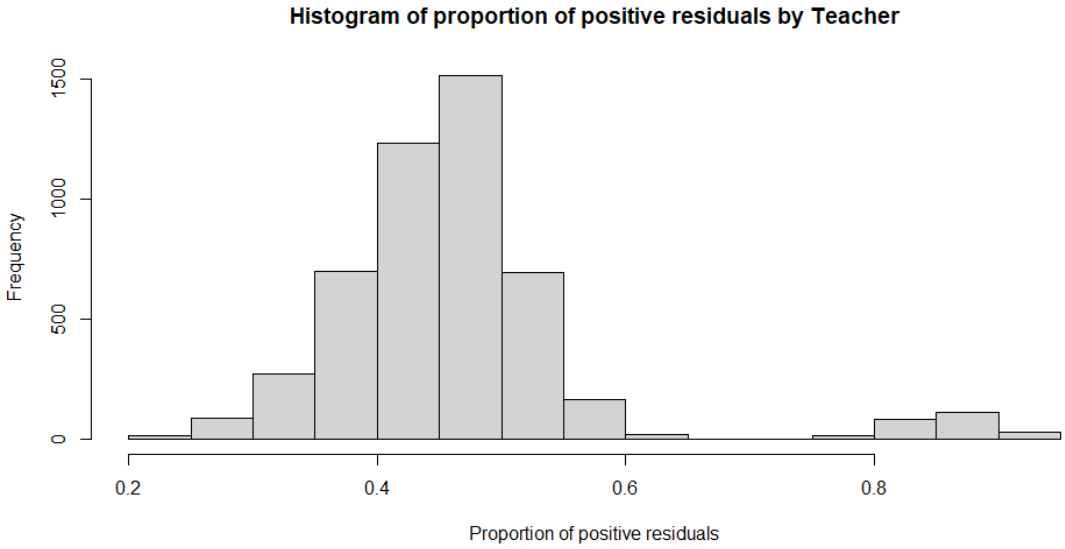
Cheater Teacher: Team K

Abstract:
This report introduces methodology to find cheater teachers from students' scores and our method to find the cheating mechanism. Teachers have a high incentive to manipulate students' scores as they are awarded if students did well in their tests. It would be reasonable to assume teachers who did manipulate students' grades will have a higher proportion of merit students in their classes. This is our underlying principle of investigating cheater teachers.

Task 1: Find the cheats:

Given their grades and their mean scores over the years, we have a general idea of students' performance in study. It is reasonable to assume that although some students may improve their grades through consistent studying, the degree of improvement should not be much different from their overall mean grade. Therefore, students who were taught by a cheater teacher would have an unreasonably high grade above their mean grade. Further, classes with cheating teachers will have a higher likelihood of having large test score fluctuations. We expect to see that deceitful teachers will have a larger proportion of merit students in their classes which distinguish them from non-cheats. We tried to investigate the distribution of positive residuals of scores, that is the positive difference between a student score and the student's mean score.

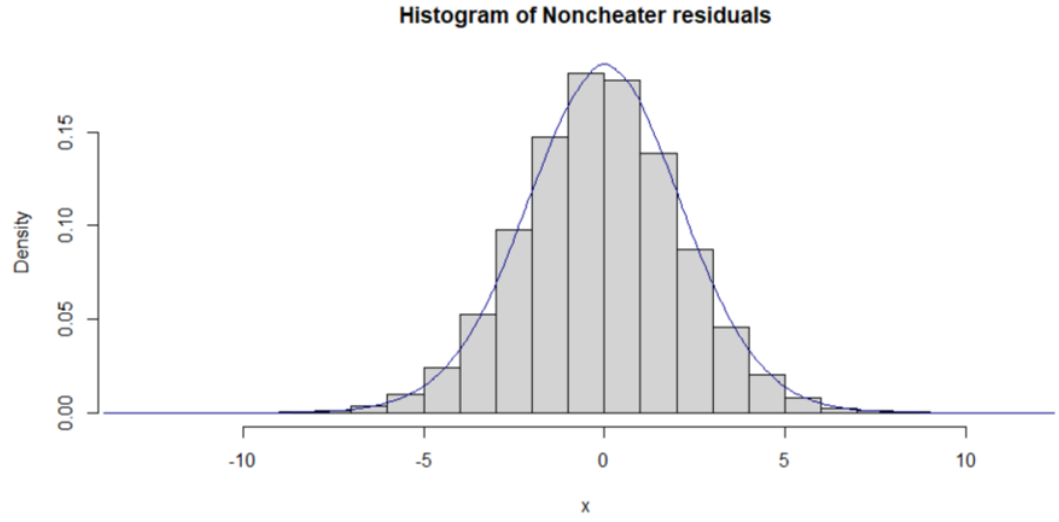


From the plot we can see that the distribution of the proportion of positive residuals is bimodal, with some unusual observations on the right, which represents teachers who were likely to manipulate students' grades. We find **74%** is the proportion that differentiate cheats from non-cheats. Hence, we filter out teachers who have a proportion of positive residuals higher than 74%, which gives **242 cheater teachers**.

Task 2: Investigating cheating mechanism:
Our aim is to learn more about the cheating mechanism used by cheating teachers. We also want to formulate a probability model to separate the students that had their grades altered by cheating teachers from those who didn't.

Step 1:

We split our data into non-cheating teachers and cheating teachers, using our list from Task 1. We then calculate the difference between each child's score and their mean score which we will call the residual. By plotting these residuals, we gain a probability density function for our non-cheating teachers.




```
noncheat_mean
  -0.222224552341218
```

```
noncheat_sd
 2.28311849956247
```

All non-cheating teachers have a normal distribution with mean -0.22 and sigma (standard deviation) 2.28.

Step 2:

We now calculate the residuals for the cheating teachers. Instead of using the MeanScore for each student we calculate their overall mean by excluding scores from cheating teachers, making our residuals more accurate.

Step 3:

Not all of the students who had cheating teachers would have had their scores altered so we want to know what proportion of students' scores were manipulated. We call this unknown parameter q.

For the proportion who did not have altered scores, their residuals should look the same as the non-cheating residuals, meaning we make a reasonable assumption that they have a normal distribution with the same parameters. We estimate that the cheating residuals also have a normal distribution but with unknown mean (mu) and sigma (std).

We therefore use Maximum Likelihood Estimation to estimate the 3 parameters; q, mu, and std.

Our function to do this (where y_i is a residual):

f(y_i) = (1-q) * dnorm(y_i, known mean, known sd) + **q * dnorm(y_i, mu, std)**

The output:

```
$minimum
[1] 81297.92

$estimate
[1] 0.7783761 7.3718424 3.2929184

$gradient
[1] -0.0019208528 -0.0003651874  0.0002563110

$code
[1] 1

$iterations
[1] 57
```

The proportion of a manipulated score for the student is 0.78, the mean is 7.37 and sigma is 3.29. (All to 2 dp).

Step 4:

We want to find the standard error and confidence interval for our 3 estimates, so we created a Bootstrap function to do this. We had to decrease the number of iterations to 100 for our bootstrap as the code ran for too long otherwise.

Output:

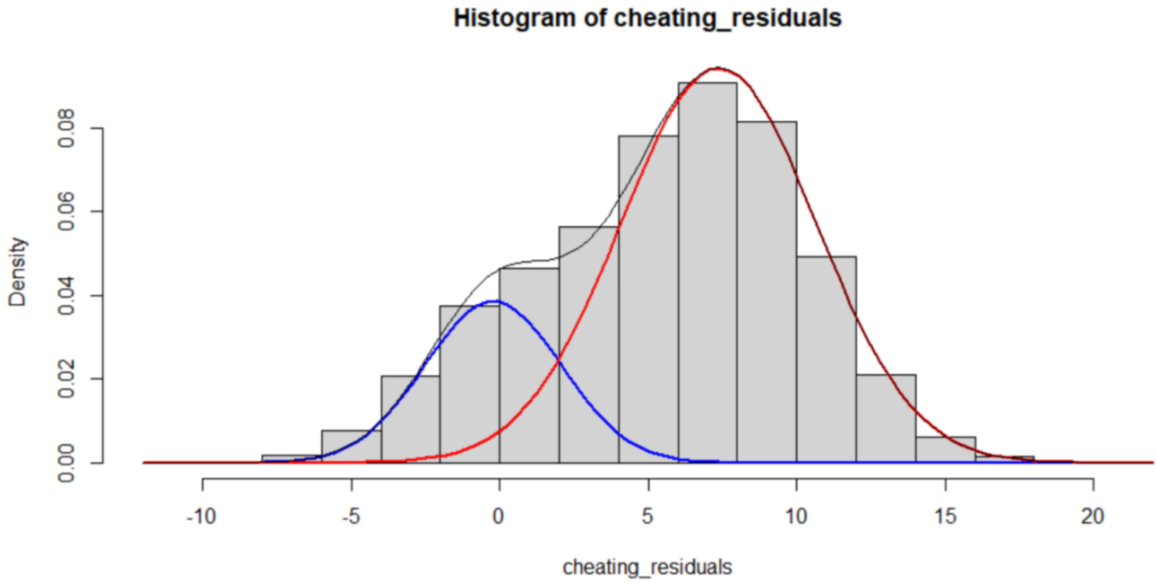
```
$stderror
     ML.q     ML.mu    ML.std
0.1306189 1.2424880 0.3191458

$CI.q.ML
     2.5%     97.5%
0.7713193 1.1826663

$CI.mu.ML
    2.5%    97.5%
3.536493 7.420144

$CI.std.ML
    2.5%    97.5%
3.254646 4.303372
```

Step 5:

To visualize our findings, we plotted the histogram of our cheating residuals and overlaid 3 lines onto it.

The black line represents the overall Probability Density Function of our data. The blue line represents the un-manipulated score residuals, which is centered at 0.

The red line represents the manipulated score residuals and describes the cheating mechanism. This distribution is centered around 7, meaning the mean residuals for cheating teachers is higher than for non-cheating teachers. This tells us on average, the manipulated score was increased by 7 points. We estimated q to be around 0.78, meaning if a student had a cheating teacher there is a 78% probability that their score was altered.



The overall probability for a student's score to be manipulated is quite large. Also, the scores that were manipulated were, on average, increased by quite a significant amount. Overall this could be enough to significantly increase a teacher's overall class grade, leading to a bonus.

Conclusion:

In summary, the cheating mechanism can be formulated in terms of a Normal distribution, with a mean of 7.37 and standard deviation of 3.29. The proportion of manipulated scores in a cheat teacher's class is 0.78. By visualising the distribution of our data, we found on average, the manipulated scores have increased by 7 points, which is a significant amount to help dishonest teachers to earn a bonus.