Assignment 1

Group 08

Due: 11:55pm Thursday 12th October 2023

Group 08

- Jason Abi Chebli (31444059)
- Sovathanak Meas (29400090)
- Farrel Wiharso (32787154)
- Neev Bhandari (32508743)

Task 1 - Fit a Distribution [30 marks]

Q1: Tidying the Data [1 Mark]

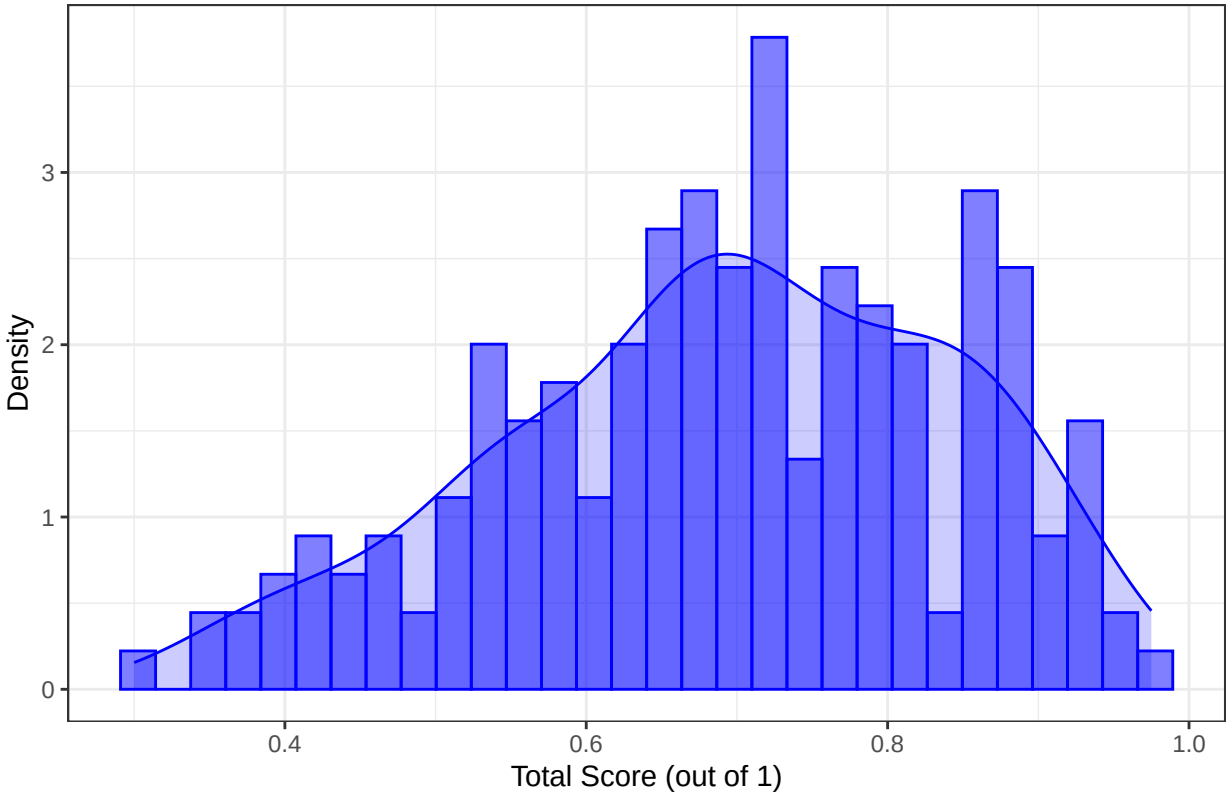
```
TidyGradesData <- GradesData |>
    filter(Total != 0)
```

The class has a total of 200 students. Out of those 200 students, only 193 gave the class quiz a genuine attempt, meaning that 7 students did not attempt the class quiz. These student's results have been omitted to avoid it from influencing our analysis.

Q2: Should we use a Normal Distribution? [3 Marks]

The distribution of the genuine attempts of the Management I class quiz can be seen in Figure 1.

Figure 1: Distribution of the Genuine Attempts of the Class Quiz



A normal distribution fits the data best if the distribution is symmetric. As can be seen in Figure 1, the data does not seem symmetrically distributed, with a longer left tail (negatively skewed). As the data is negatively skewed and not symmetric, the normal distribution is not ideal. It would produce somewhat accurate results, but it is not the best fit model available given its differing characteristics.

Q3: Should we use a Beta Distribution ? [2 Marks]

The data is bounded from (0,1] as the maximum score is 1 and we cleared up any "non-attempts" (scores that are 0). Hence, as the values are within (0,1] and the data is negatively skewed, a beta distribution may be a valid alternative in this case, as it covers a wide range of shapes, depending on the values of α and β. It can indeed take into account the negative skewness in its parameters and, therefore, be a more appropriate fit than the normal distribution. Furthermore, the beta distribution captures the distribution's unpredictable variance with better effectiveness.

Q4: Normal vs Beta Distribution [8 Marks]

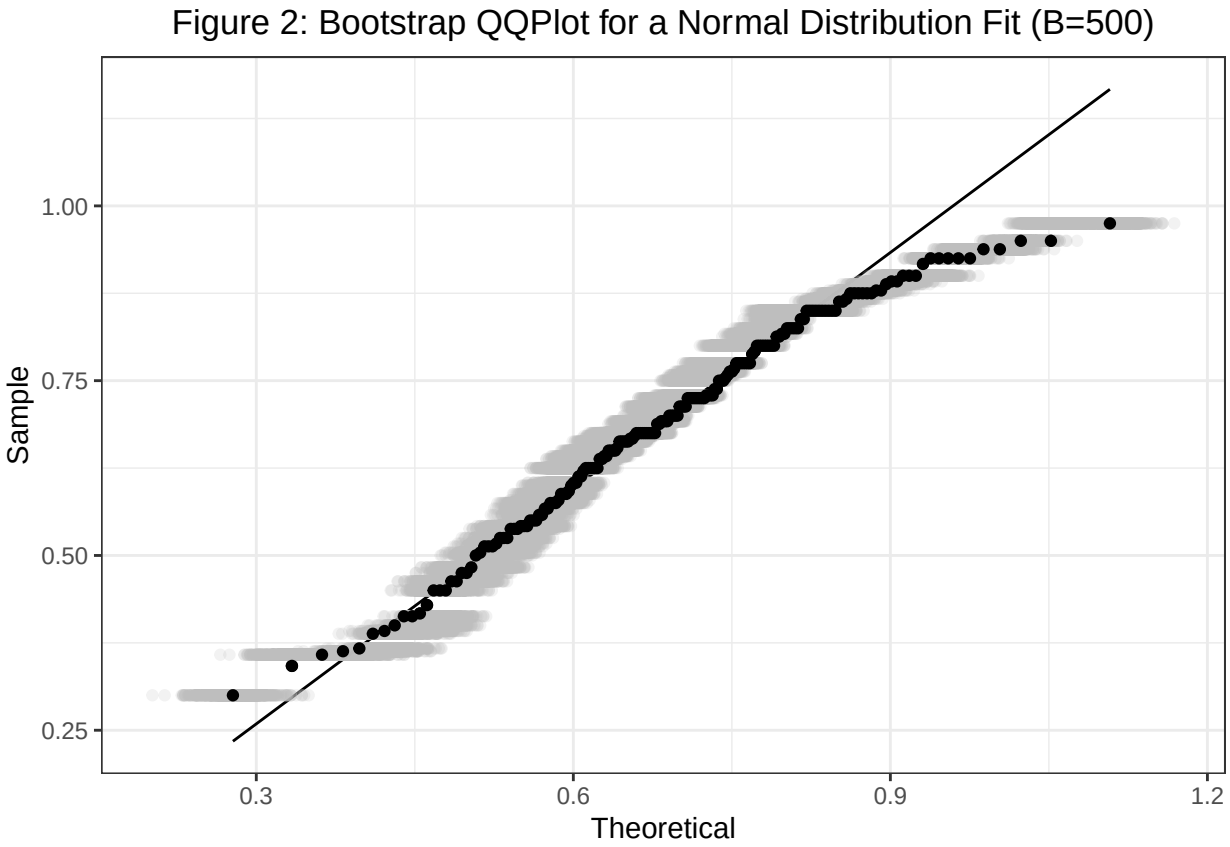
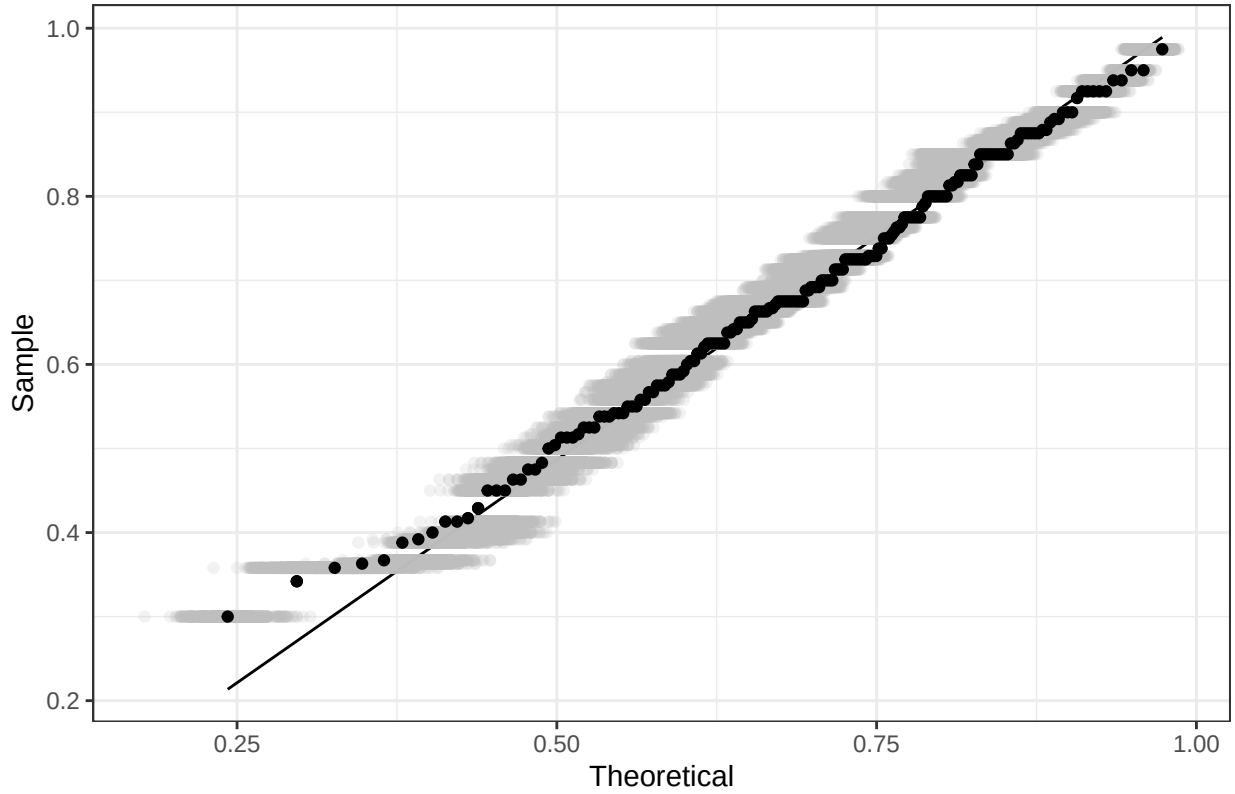


Figure 3: Bootstrap QQPlot for a Beta Distribution Fit (B=500)



Fitting a normal distribution to the data, we were able to determine the MLE estimates for the parameters, $(\hat{\mu}, \hat{\sigma})$, given that there are two population parameters. As such,

$$\theta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}, \quad \hat{\theta}_{MLE} = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = \begin{pmatrix} 0.693 \\ 0.148 \end{pmatrix}$$

The 95% Bootstrap Confidence Interval's for μ are 0.6733123 and 0.715928.

The 95% Bootstrap Confidence Interval's for σ are 0.1349123 and 0.1605527.

Fitting a beta distribution to the data, we were able to determine the MLE estimates for the parameters, $(\hat{\alpha}, \hat{\beta})$, given that there are two population parameters. As such,

$$\theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \hat{\theta}_{MLE} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} 5.95 \\ 2.63 \end{pmatrix}$$

The 95% Bootstrap Confidence Interval's for α are 5.0425924 and 7.2263702.

The 95% Bootstrap Confidence Interval's for β are 2.2094739 and 3.1367506.

From the QQPlots, we would recommend a beta distribution as it is a better fit. To come to this conclusion, a 'thick-marker' judgment approach was used. A distribution is a good fit if all of the values line-up on the 45 degree line. Furthermore, we can use a 'thick-marker' to draw over the line, where a good fit should cover most if not all the points. This was done over the qqplots in Figure 2 and Figure 3, to generate Figure 4 and Figure 5.

Figure 4: QQplot (Normal Dist)

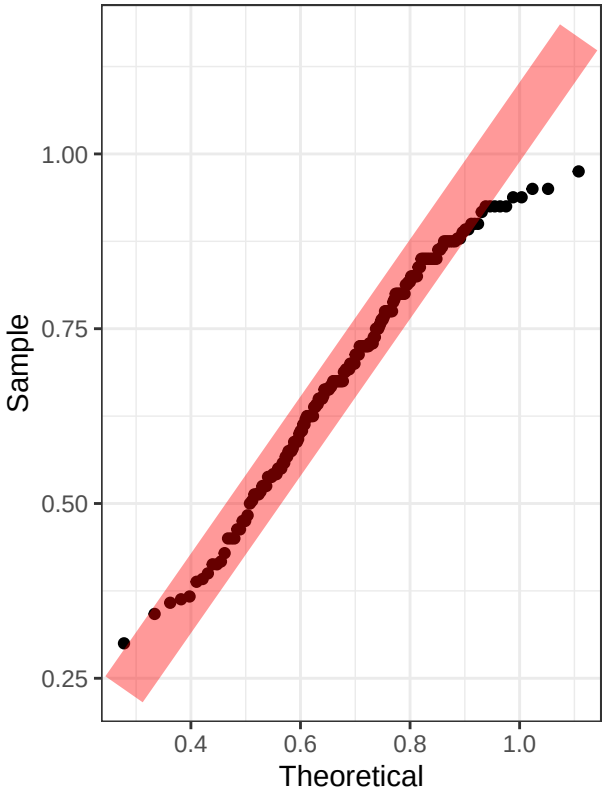
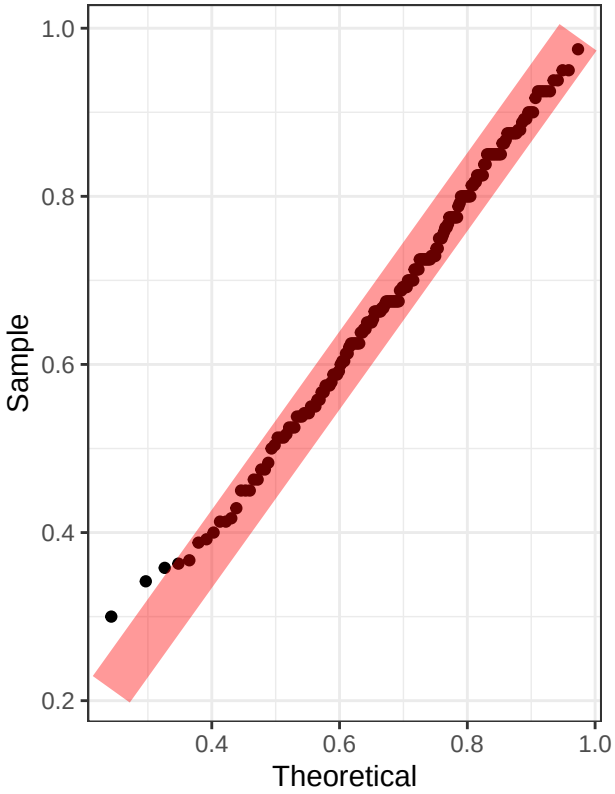


Figure 5: QQplot (Beta Dist)

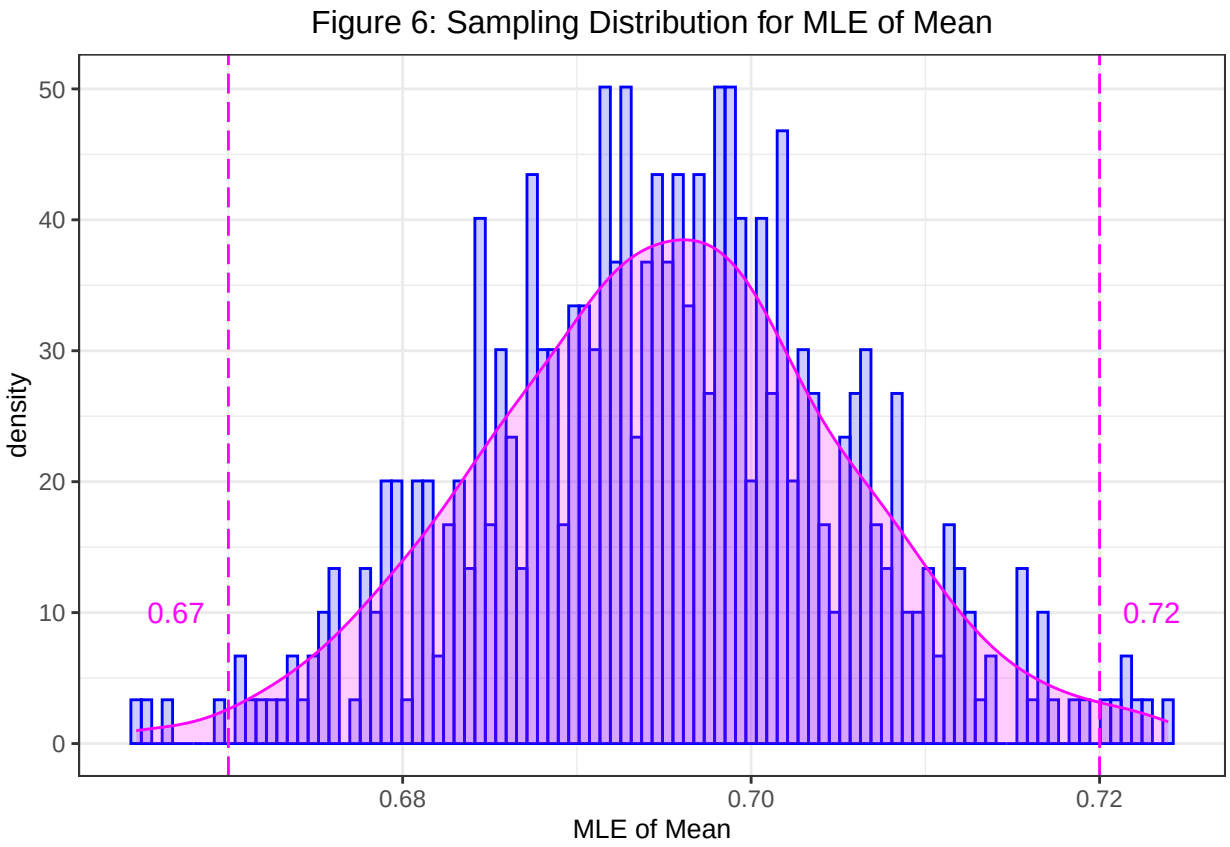


As seen in Figure 4 and Figure 5, if we applied a thick marker line to the qqplot of the normal distribution, we may have some points not covered at the top and maybe at the bottom. Meanwhile, if we applied a thick marker line approach to the qqplot of the beta distribution, there may be less points which are uncovered, with only a few at the bottom as seen in Figure 5. Hence, comparing Figure 4 and 5 illustrates that the beta distribution is a better fit, thus recommending its implementation in this scenario over a normal distribution.

Q5: Mean and Median of the Grade Distribution [4 Marks]

Using a beta distribution fit, we estimate the population mean to be 0.6935934 and the population median to be 0.7092012. We can interpret these numbers as the following. The average score on the quiz is 0.6935934. Meanwhile, 50% of students got a score about 0.7092012 and 50% of students got a score below 0.7092012 on the quiz.

Q6: Plot and interpret a 99% parametric bootstrap [4 marks]



We are 99% confident that the parametric bootstrap of the mean for the beta distribution lies between 0.6679346 and 0.7220877. From Figure 4, the lecturers goal of the average quiz mark being 0.7 lies within this range and so it is possible that this aim was achieved through the testing.

Q7: Average, Proportion Failed, Proportion of HDs [4 marks]

The average grade is 69.3593445%.

The estimated proportion of students within 15% of this average is 0.4865063.

Estimated number of students within 15% of the average: 93.8957236.

The estimated proportion of students that would fail (a score below 60%) is 0.2583398.

Estimated number of students who failed: 49.8595875.

The estimated proportion of students that would get HD (a score above 80%) is 0.2680998.

Estimated number of students with HDs: 51.7432553.

Q8: Do you think that the quiz achieved the lecturer's aims? [4 marks]

Overall, we can conclude that the aim of the quiz was partially achieved. The average was approximately similar to the target of 70%. However, the majority of students did not lie within the +- 10 % of the mean score, and a considerable proportion failed to achieve the passing mark of 60%. As a result, the variance in quiz scores is perhaps higher than expected, which has consequently led to these outcomes.

Task 2 - Are Post Grad Students better? [30 marks]

Note: Task 2 continues to utilise the cleaned data.

Q1: New Variable to Data - Pass or Fail [4 marks]

```
TidyGradesDataWithResults <- TidyGradesData |>
    mutate(Result = ifelse(Total >= 0.6, "PASS", "FAIL"))
```

Q2: Relevant Proportion Table [10 marks]

Table 1 outlines the relevant proportions of passes and fails for the different cohorts.

Table 1: Proportion of Fail and Pass Results by Cohort

Cohort	Fail Proportion	Pass Proportion
PG	0.146	0.854
UG	0.297	0.703

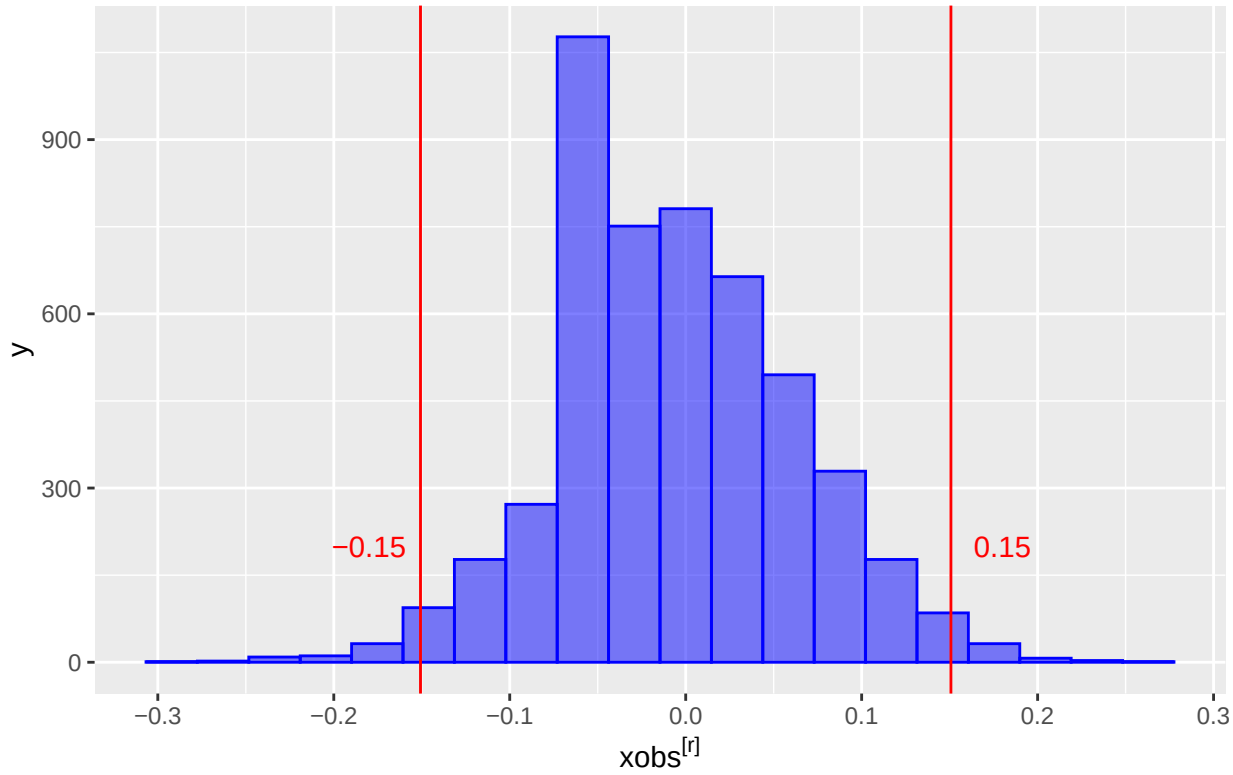
As can be seen in Table 1, 85.4% of Post Graduate students pass the quiz, while 70.3% of Under Graduate students pass the quiz. This indicates that, regarding this quiz, Post Graduate students perform better than undergraduate students, with around 2.03 times more undergraduate students failing than post graduate students.

At the same time however, it is noticeable that there are significantly more undergraduate students taking the quiz compared to postgraduates. As a result, to convert this into meaningful analysis, we will need to investigate further, such as utilizing permutation tests and hypothesis testing.

Q3: Permutation Test [4 marks]

The p-value obtained from the permutation test is: 0.0554.

Figure 7: Approximate Sampling Distribution of $\hat{p}_{PG} - \hat{p}_{UG}$ under H_0



Q4: What is the result of this test? [8 marks]

To determine whether postgraduate students are better, we want to see if the proportion of postgraduate students that pass the quiz, denoted as $\hat{p}_{PG}$, is **statistically significantly** different to the proportion of undergraduate students that pass the quiz, denoted as $\hat{p}_{UG}$. So, we test:

$H_0 : \hat{p}_{PG} - \hat{p}_{UG} = 0$

$H_1 : \hat{p}_{PG} - \hat{p}_{UG} \neq 0$

or if we let $\delta = \hat{p}_{PG} - \hat{p}_{UG}$, then our null and alternative hypothesis are:

$H_0 : \delta = 0$

$H_1 : \delta \neq 0$

After conducting a permutation test with 5000 replications, the distribution is shown in Figure 7. The plot in Figure 7 represents a sampling distribution of proportion differences generated under the assumption that H0 is true (i.e.that postgraduate students are the same as undergraduate students). The red line shows the proportion difference that we observed in our sample.

As can be seen, the sampling distribution is approximately normally distributed around a mean value of 0 and our observation in our sample seems to be along the right tail. What this signifies is that our observation may simply be a rare case.

The p-value from the randomization test is 0.0554. This indicates that at a 5% and 1% significance level, there is no statistically significant difference and so we are in favor of the null hypothesis as we cannot reject the claim that post graduate students are the same as undergraduate students (H_0).

However, at a 10% significance level, there is a statistically significant difference and so we are in favor of the alternative hypothesis as we can reject the claim that post graduate students are the same as undergraduate students (H_0). Based on the evidence we have (i.e. our sample data and the observed difference in proportion), it is unlikely that we would have observed a difference this large by chance. We can see this visually from the plot. Our observed difference (the red line) is not likely to have come from a distribution which assumes that there is no difference in how post graduates perform compared to undergraduates, since the red line is far from 0.

Q5: Why p-value is not 0? [4 marks]

Such a large difference in proportions does not have a p-value of almost zero. This may occur for multiple reasons including:

1. Undergraduate students and post graduate students are not two very separated cohorts. They are still all taught the same content and assessed in the similar manner. As such, they share a lot of similarity, meaning that there may not be such a large discrepancy between them and hence why the p-value is not exactly 0 as the larger the discrepancy the smaller the p-value would be.
2. Despite the large difference in the passing rates of the two samples, there is a significant disparity between the UG and PG sample sizes, with there being almost as three times more undergraduates taking the quiz. As a result in order to reduce the p-value further, stronger evidence is required in the postgraduate cohort, which could be from an increased sample size.
3. Due to the sample size difference, the variance within each of the groups also differs which would cause the p-value to be higher than expected. As the p-value considers the variability of each of the groups, when one of the groups have a higher variability, in this case, it can be expected that the data of the postgraduates have higher variance due to the smaller sample size which explains the high p-value.

Task 3 - Bayesian Analysis [33 marks]

Note: Task 3 continues to utilise the cleaned data.

Q1: Prior Distributions [2 marks]

The proportion of undergraduate students and post graduate students that passed can be seen in Figure 8 and 9 respectively.

Figure 8: Distribution of Undergraduate Passing Proportion

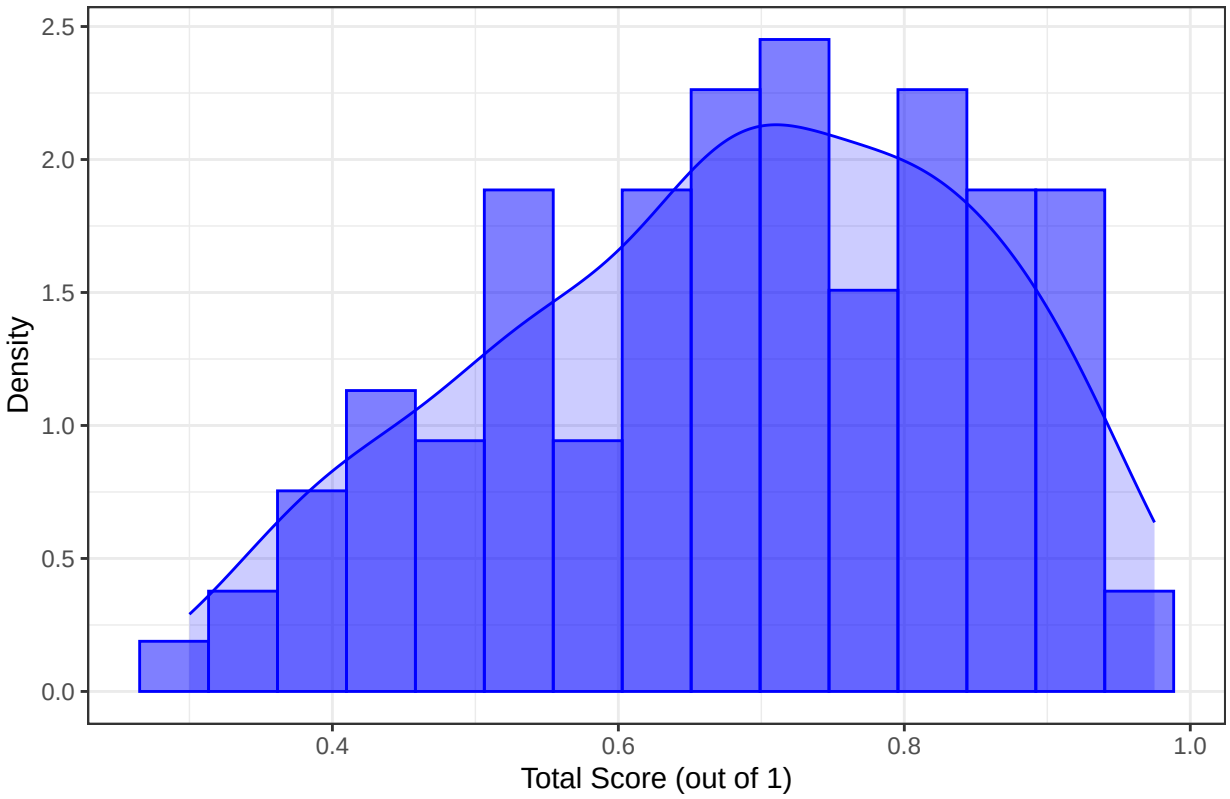


Figure 9: Distribution of Postgraduate Passing Proportion

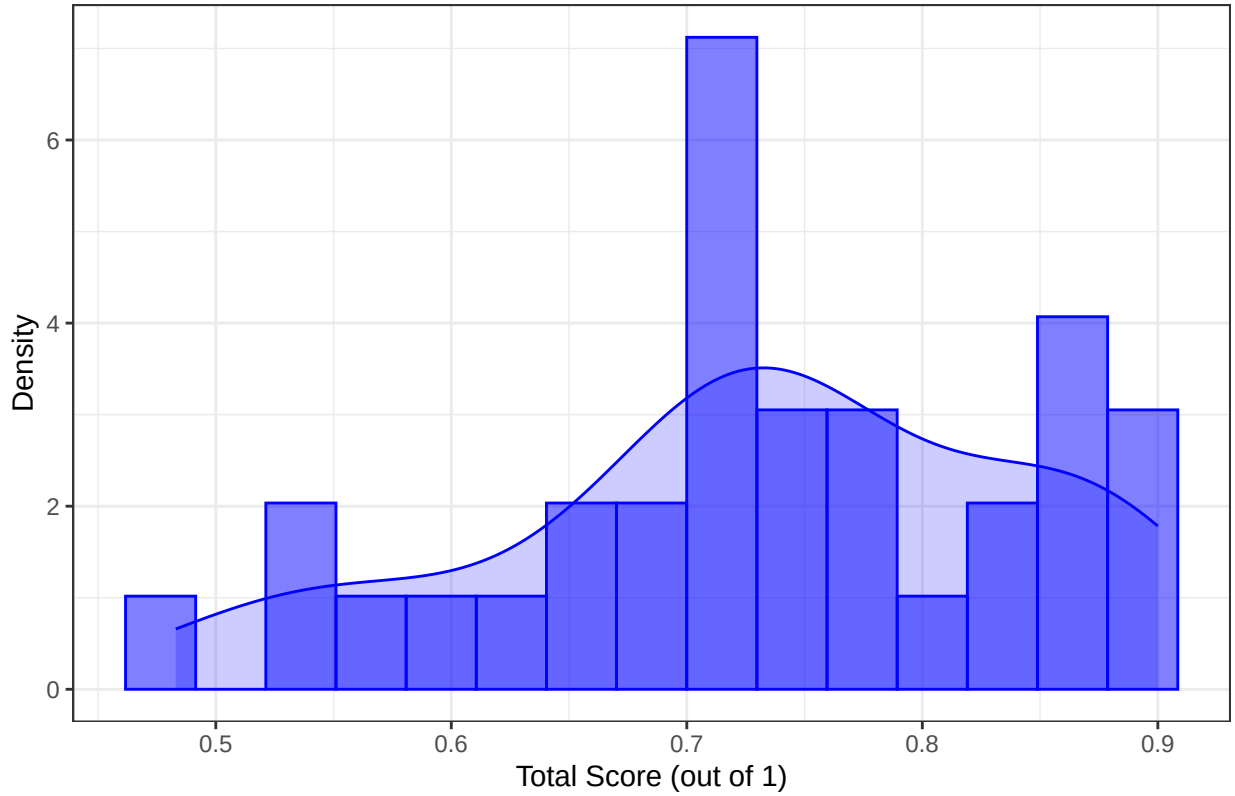


Figure 8 illustrates the distribution for undergraduate students that "passed" the quiz. As can be seen, the distribution seems to be a continuous bi-modal distribution. To an extent, it looks similar to a continuous uniform distribution and as such, an appropriate prior distribution for the proportion of undergraduate students who "passed" could be a beta distribution as this distribution can take many shapes, including a skewed or somewhat uniform distribution.

Figure 9, on the other hand, illustrates the distribution for postgraduate students that "passed" the quiz. Similarly, it seems to also be a continuous bi-modal distribution. However, the distribution is not as uniform and could be argued to be slightly left skewed. Similarly, a beta distribution could also be used to fit the data, as beta distributions can take into account this skewness. As such, an appropriate prior distribution for the proportion of postgraduate students who "passed" could be a beta distribution.

Please note that although both seem bi-modal, there is no distribution that can appropriately 100% fit a bi-modal model. As such, a uni-modal model must be chosen. Furthermore, given that we do not yet know what alpha and beta to use in the prior beta distribution, the prior distribution for both will be: $beta(1, 1)$

Q2: Posterior Distributions [2 marks]

To determine the posterior distributions for the proportion of undergraduate students who "passed" and one for the proportion of postgraduate students who "passed", we will use the prior's conjugate pair.

For the proportion of undergraduate students and postgraduate students that "passed", the prior distribution was a beta distribution. This means that the posterior can either be Bernoulli or Binomial, as these are the only two conjugate pairs with beta. Bernoulli is a discrete distribution and Binomial is a continuous distribution. Therefore, given that the proportion is a continuous distribution, the posterior distribution for the proportion of undergraduate and postgraduate students who "passed" must be a binomial distribution.

In Summary, the appropriate prior-posterior distribution chosen are:

- Proportion of undergraduate students who "passed": Beta-Binomial
- Proportion of postgraduate students who "passed": Beta-Binomial

Q3: 95% Credibility Intervals [4 marks]

There is a 95% probability that the true value of the mean of the postgraduate students that passed lies within the interval [0.7276,0.9268], based on the data and our prior beliefs.

There is a 95% probability that the true value of the mean of the undergraduate students that passed lies within the interval [0.6245,0.7717], based on the data and our prior beliefs.

As can be seen, for the postgraduate students, the interval is larger than that for the undergraduate students. This indicates greater disparity and more uncertainty/variability with the postgraduate students mean, than the undergraduate students mean. This is likely to be attributed to the lower sample size representing this cohort, thus reducing the confidence in the estimate's accuracy due to lesser data being available.

Q4: The estimator that minimises the posterior expected squared error loss [8 marks]

The estimator is in a form that linearly combines a purely data-based estimator and some prior quantity. In fact, the form of the estimator for a beta binomial is:

$$\hat{\mu} = Z \times (x/n) + (1 - Z) \times \mu_{prior}$$

This form is true for both the undergraduate students and post graduate students as they are both a beta-binomial conjugate pair.

For the undergraduate cohort:

Estimate ($\hat{\mu}_{UG}$) = 0.7007

Credibility Factor (Z) = 0.9864

For the post graduate cohort:

Estimate ($\hat{\mu}_{PG}$) = 0.84

Credibility Factor (Z) = 0.96

As can be seen, the credibility factor for both of the cohorts is quite high (the maximum credibility factor value is 1). As such, this implies that the prior distribution is heavily relied upon compared to the likelihood distribution. This is not the best outcome as we made an educated "guess" of the prior. Instead, we would prefer it if the credibility factor was lower, as it would indicate that our posterior distribution is not heavily influenced by our "guess" of the prior distribution.

The estimates that minimize the posterior expected squared error loss are 0.7007 and 0.84 for the undergraduate and postgraduate cohort, respectively. These estimates are the mean for the respective cohorts that will result in the most appropriate fit and minimizes the expected squared error loss function:

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

Q5: The estimator that minimises the posterior expected absolute error loss [4 marks]

The estimator that minimises the posterior expected absolute error loss for the proportion of students who "passed" for a given cohort is the median.

Hence,

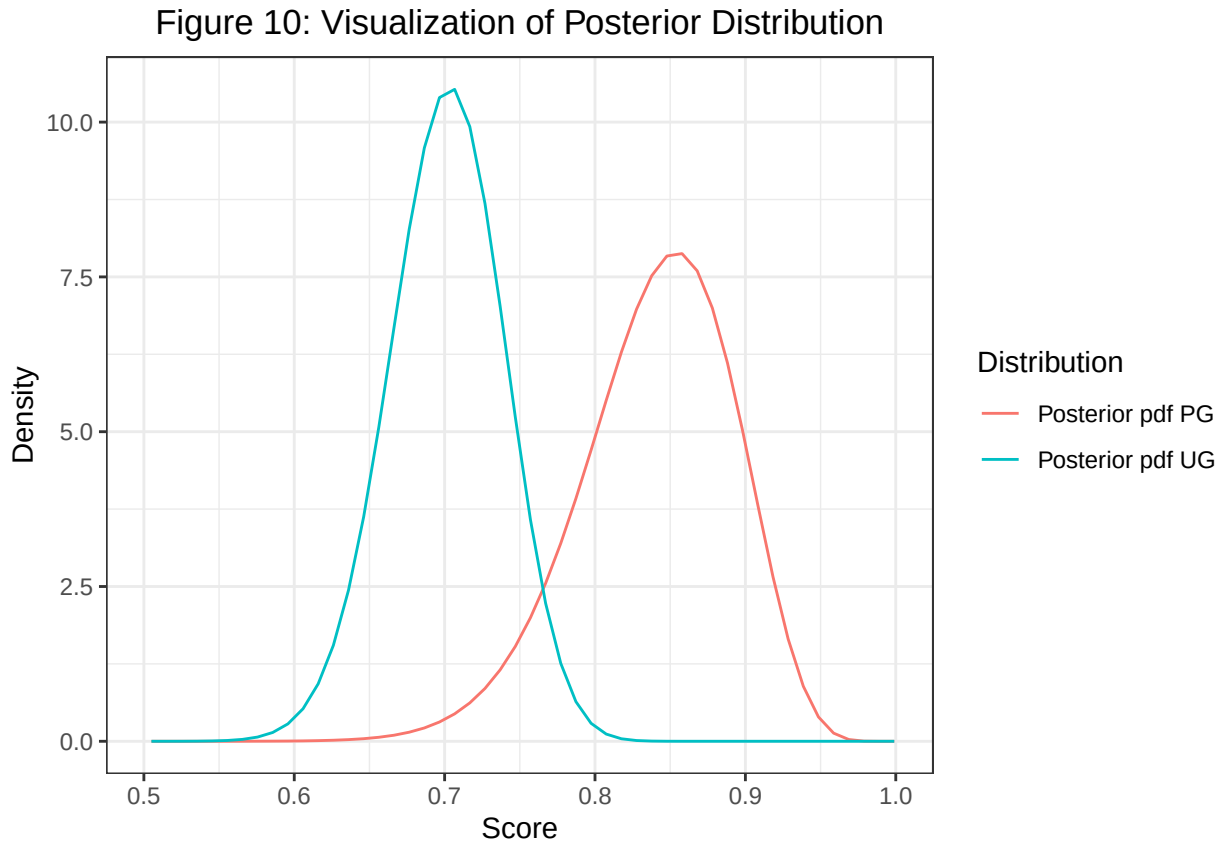
The estimate for for the undergraduate cohort: 0.7016

The estimate for for the post graduate cohort: 0.8445

The estimates that minimize the posterior expected absolute error loss are 0.7016 and 0.8445 for the undergraduate and postgraduate cohort, respectively. These estimates are the mean for the respective cohorts that will result in the most appropriate fit and minimizes the expected absolute error loss function:

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

Q6: Visualise the posterior distribution [3 marks]



As can be seen in Figure 10, the posterior mean is higher for the postgraduate students compared to that of the undergraduate students. This indicates that we expect post graduate students to perform better on average. Additionally, it is interesting to note the different densities of the two distributions. The undergraduate PDF distribution is more narrow and taller, potentially indicating that the performance of undergraduate students is more consistent, compared to that of post graduate students who have a shorter and slightly wider posterior pdf distribution.

Q7: Which prior to use in future [2 marks]

Even though our prior should (in theory) not have much of an influence on our data (given a large enough sample), it is still important to update our beliefs. Given that we now know the posterior distribution from this analysis, in the future, the prior distribution, should be the current posterior distribution. This means that for both cohorts, the prior should be a beta distribution $X \sim beta(\alpha, \beta)$. Overall, this is optimal since the grade and pass distributions obtained from this semester's data serve as a strong reference point for future cohorts to compare against.

Specifically,

for the undergraduate cohort: $X \sim beta(103, 44)$, and

for the postgraduate cohort: $X \sim beta(42, 8)$.

Q8: The posterior distribution of the difference in the proportion [4 marks]

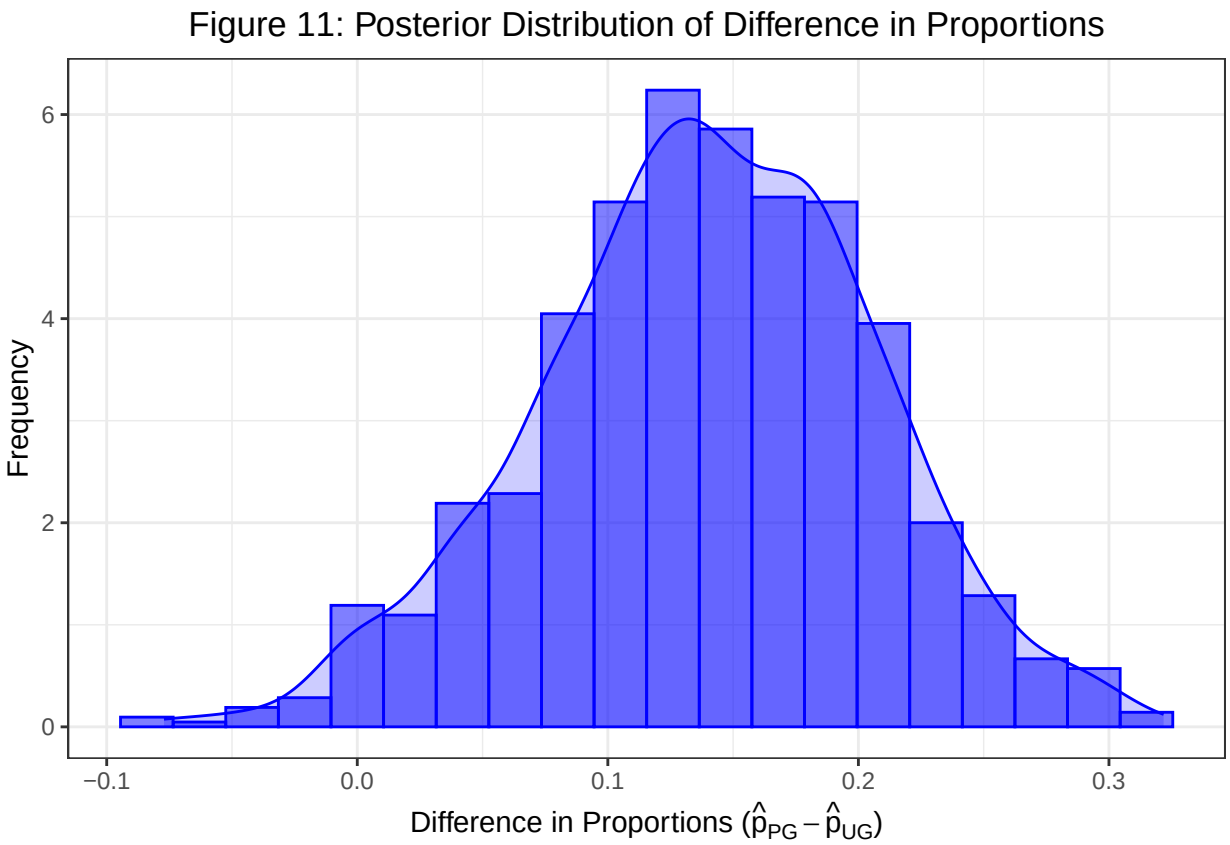
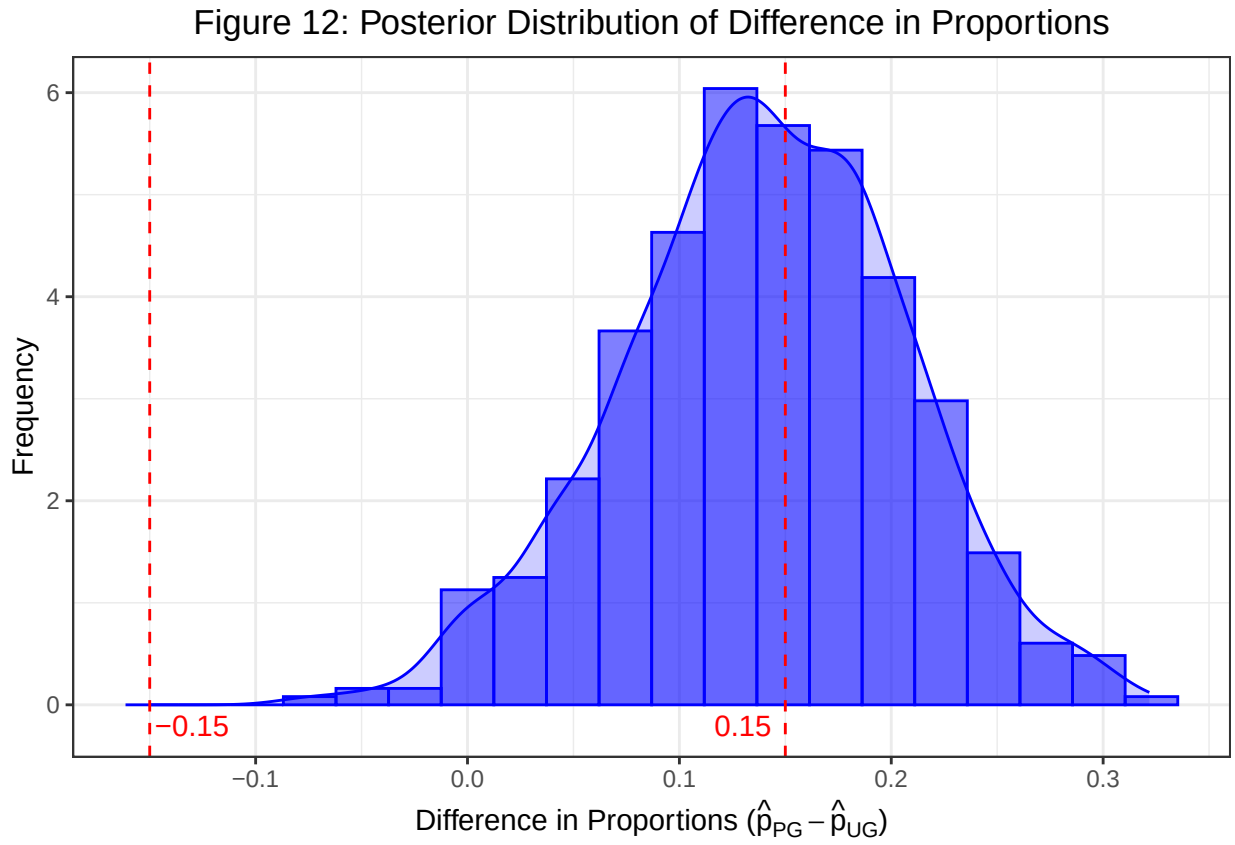


Figure 11 shows that the posterior distribution of the difference in proportions is somewhat normally distributed (slightly negatively skewed). This distribution is not around 0.0, indicating that the average proportion of postgraduate students that passed is greater than the proportion of undergraduate students that passed, suggesting that post graduate students perform better on average.

Q9: The probability that the difference in the grades of the 2 cohorts is within 0.15 [4 marks]



The probability that the difference in the grades of the 2 cohorts is within 0.15 is 0.548. As can be seen in Figure 12, this value makes sense as it looks like roughly half of the values lie within the -0.15 to 0.15 range.