Group16_A1

Alexandra Goh (29796431), Vaishnavi Maganti (33181837), Lucy Field (32527284), Shalini Nair (33696411)

2023-09-17

1 Fit a distribution

1.1 Checking and modifying the data

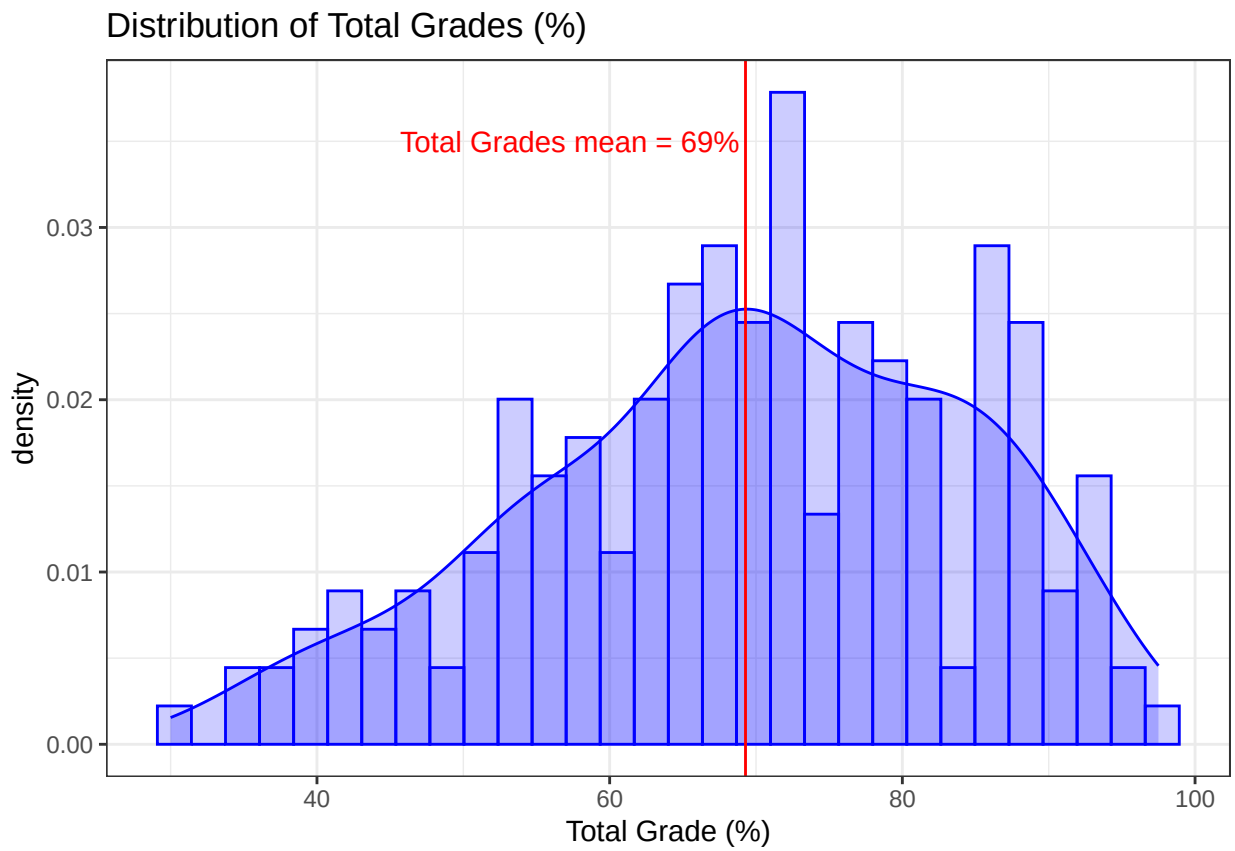
We have removed all zero attempts for each question, as seen in the code below.

```
grades <- read_csv("GradesData.csv")
grades <- subset(grades, Total != 0)
```

1.2 Is our data normally distributed?

In this scenario, it is not a good idea to use a normal distribution to model grade distributions. The reason being, based on the density plot and histogram constructed below, the data appears to be multimodal, meaning it has multiple peaks and is not following a typical normal distribution pattern. A normal distribution is characterized by a single central peak and a symmetric bell-shaped curve.

Additionally, the grade distribution seems to exhibit some negative skewness. This implies that the data is not symmetric and is stretched out towards the higher grades, suggesting that a significant proportion of students may have higher performance levels. Therefore, trying to model this data with a normal distribution would not accurately capture the observed skewness.



1.3 Using a Beta distribution

The Beta distribution is appropriate for this case as it is a bounded continuous distribution. This means that it is defined on a finite interval (i.e. only taking on values between 0 and 1), therefore making it suitable for modelling grade distributions as it is not possible to have grades below 0% or over 100%. Moreover, the Beta distribution provides a continuous probability density function which is essential for modeling continuous data like grade percentages. It is also often used to model the uncertainty about proportions or probabilities, which aligns well with representing grades as proportions of a maximum possible score (e.g., achieving 80% out of 100%). Last but not least, the Beta distribution can account for skewness in the data. While the Normal distribution assumes perfect symmetry, the Beta distribution can accommodate situations where the grade distribution is not symmetric; for instance, if there is a higher proportion of students with grades closer to 100% and fewer students with lower grades, the Beta distribution can capture this skewness more effectively.

1.4 Using Maximum Likelihood to determine which distribution fits best

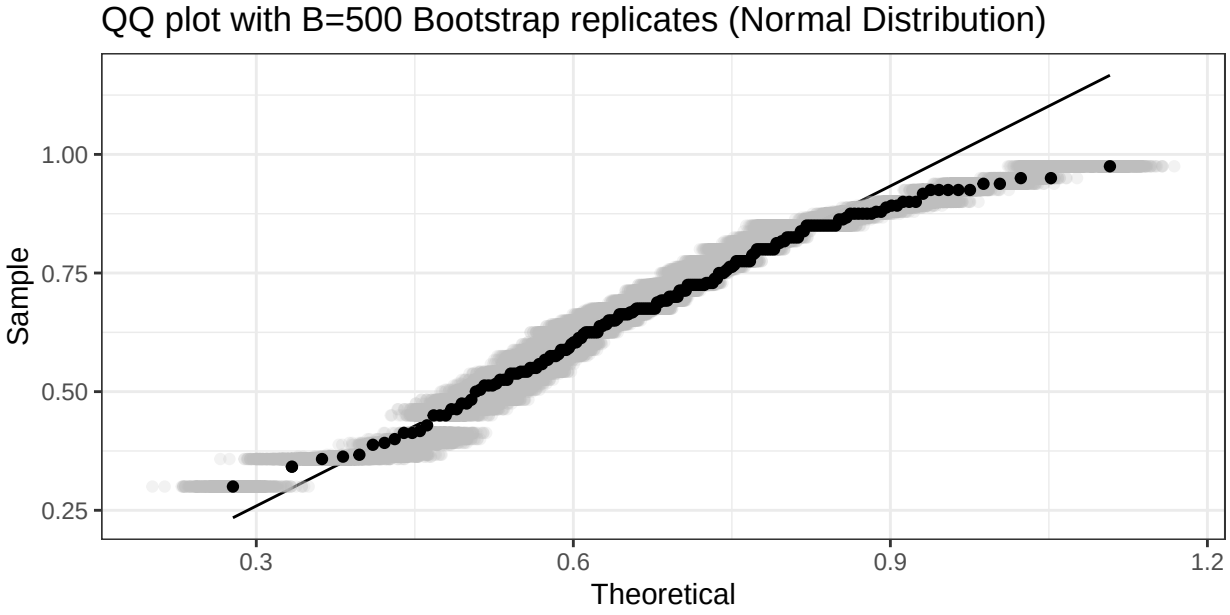
We would recommend the Beta distribution. Based on the bootstrap QQplots generated, the QQplot for the Beta distribution shows that most of the points form a roughly straight line along the reference line (with only slight deviations). Some data points in the lower tail values do deviate from the reference line, but this may just be due to randomness and can be considered as minor outliers.

From the Normal distribution's QQplot, it appears that the Normal distribution struggles in capturing the behavior of lower and upper values in the dataset. This is particularly evident in the large tails, especially at the upper end of the distribution. This suggests a potential mismatch between the observed data and the Normal distribution, as large tails imply that the Normal distribution may not adequately account for

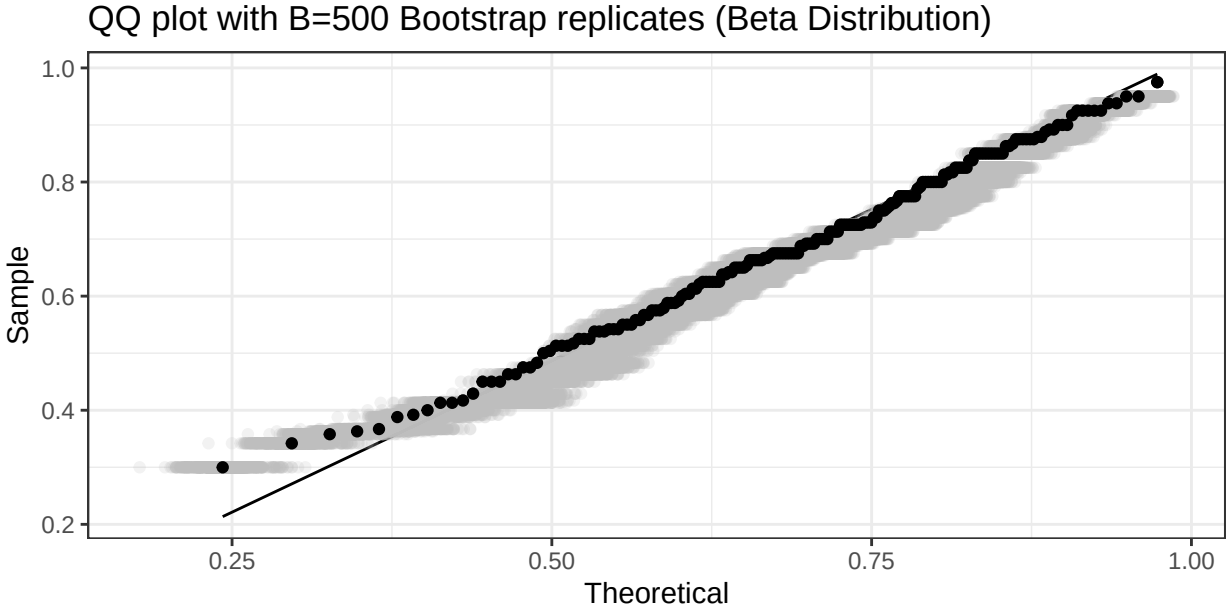
extreme values in the dataset, which might be important (e.g. if the lecturer wants to identify students who excelled or struggled significantly with the quiz).

Overall, the Beta distribution seems to provide a better overall fit to the data, as the points in its QQplot are better-aligned with the reference line relative to the Normal distribution.

Normal Distribution



Beta Distribution



1.5 Mean and Median of the grade distribution

```
MLE.x <- betadist_fit$estimate
MLE_SE.x <- betadist_fit$sd
```

Using our derived MLE's, the estimated value of alpha (α) is 5.95 whereas the estimated value of beta (β) corresponds to 2.63. To find the mean of the Beta distribution (μ), we can calculate it as:

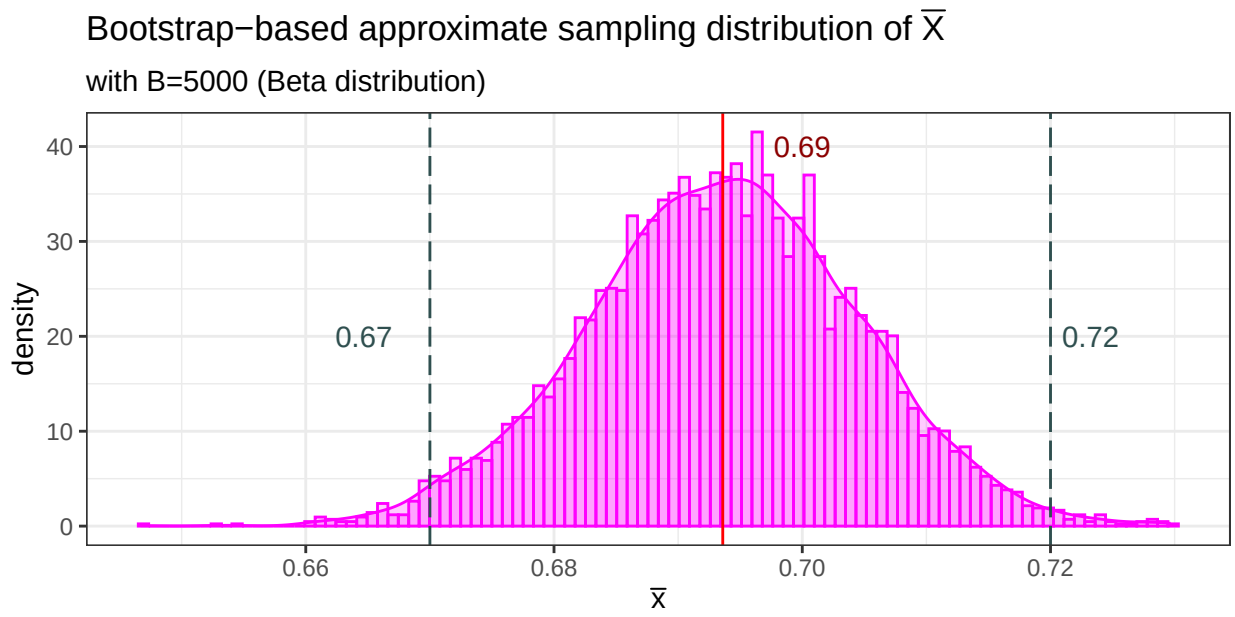
$$\mu = \frac{\alpha}{\alpha + \beta}$$

Therefore, the mean of the grade distribution is: 0.694. This suggests that the average mark, according to our Beta distribution model, is estimated to be 69.4% which almost aligns with the lecturer's hope for the average mark to be around 70%. Using the `qbeta` function, the median is calculated to be 0.709; this means that approximately 50% of students are expected to score less than or equal to 71%, and the remaining 50% are expected to score greater than 71%.

$$\text{Mean} = \frac{5.95}{5.95 + 2.63} = 0.694$$

```
qbeta(0.5, shape1 = betadist_fit$estimate[1], shape2 = betadist_fit$estimate[2])
```

1.6 99% Parametric Bootstrap of the Mean (Beta Distribution)



According to our parametric bootstrap analysis of the mean of the Beta distribution, the 99% bootstrap confidence interval for the average quiz mark ($\bar{x}$) are 0.667 and 0.722. This is a relatively narrow interval, as we're 99% confident that the true mean of the quiz marks lies between 67% and 72%.

Considering that the lecturer's goal was to achieve an average mark of around 70%, we can conclude that our bootstrap confidence interval for the mean does include this value. Hence, based on this parametric bootstrap analysis, the average quiz mark is consistent with the lecturer's goal. The interval of 67% to 72% is within close proximity to the targeted goal of approximately 70%, suggesting that the lecturer's aim has been supported by the data.

1.7 Proportion Analysis for Student Grades

Using the MLE's, we estimated the approximate proportion of students that will score within 15% of the average mark is 0.49 (i.e. 49%). According to the lecturer's passing benchmark of 60%, the proportion of students that would have failed the quiz is 0.26 (i.e. 26%). Meanwhile, assuming that a HD grade is 80%, we estimated that the proportion of students to score a HD would be 0.27 (i.e. 27%).

```
# Define the range within 15% of the mean

lower_bound <- xbar.x - (0.15 * xbar.x)
upper_bound <- xbar.x + (0.15 * xbar.x)

prop_15 <- pbeta(upper_bound, shape1 = shape1, shape2 = shape2) -
    pbeta(lower_bound, shape1 = shape1, shape2 = shape2)

fail <- 0.6
prop_fail <- pbeta(fail, shape1 = shape1, shape2 = shape2)

HD <- 0.8
prop_HD <- 1 - pbeta(HD, shape1 = shape1, shape2 = shape2)
```

1.8 Did the quiz achieve the lecturers aim's?

In summary, we believe our analysis supports the view that the quiz successfully achieved the lecturer's goals. Firstly, our MLE of the average mark, according to the Beta distribution, closely approximates the lecturer's target at 69.4% (0.694), aligning well with the lecturer's goal of hoping for the average mark to be around 70%. Furthermore, based on our constructed 99% confidence interval from parametric bootstrapping of the Beta distribution's mean, we're 99% confident that the true mean of the quiz marks lies between 67% and 72%. This narrow range indicates a high level of confidence in the mean estimate's accuracy. Last but not least, our MLE indicates that roughly 26% of students would fail the quiz. Considering the lecturer's aim to minimize the number of students scoring below the passing mark of 60%, this proportion remains within an acceptable range.

2 Are Postgrad students better?

2.1 Creating a "Result" variable

We created a new variable known as `Result`, which contains the values "Pass" or "Fail" depending on the value of the variable `Total`, and added it to our dataframe.

```
grades <- grades %>%
    mutate(Result = ifelse(Total >= 0.6, "Pass", "Fail"))
```

2.2 Table of Proportions (Students Grade Performance)

Table 1 presents a comprehensive overview of student performance categorized by "Result" and "Cohort." It separates students into two primary groups: those who "Fail" and those who "Pass," based on the achievement of a passing grade of 60% or higher. Additionally, the table distinguishes between "UG" (undergraduate) and "PG" (postgraduate) students, facilitating a comparison between these cohorts.

The "Total Students" column reveals the exact number of students within each subgroup. For instance, it indicates that 7 postgraduate students did not meet the passing criteria, while a substantial 102 undergraduate

Table 1: Summary of Student Grades by Result and Cohort

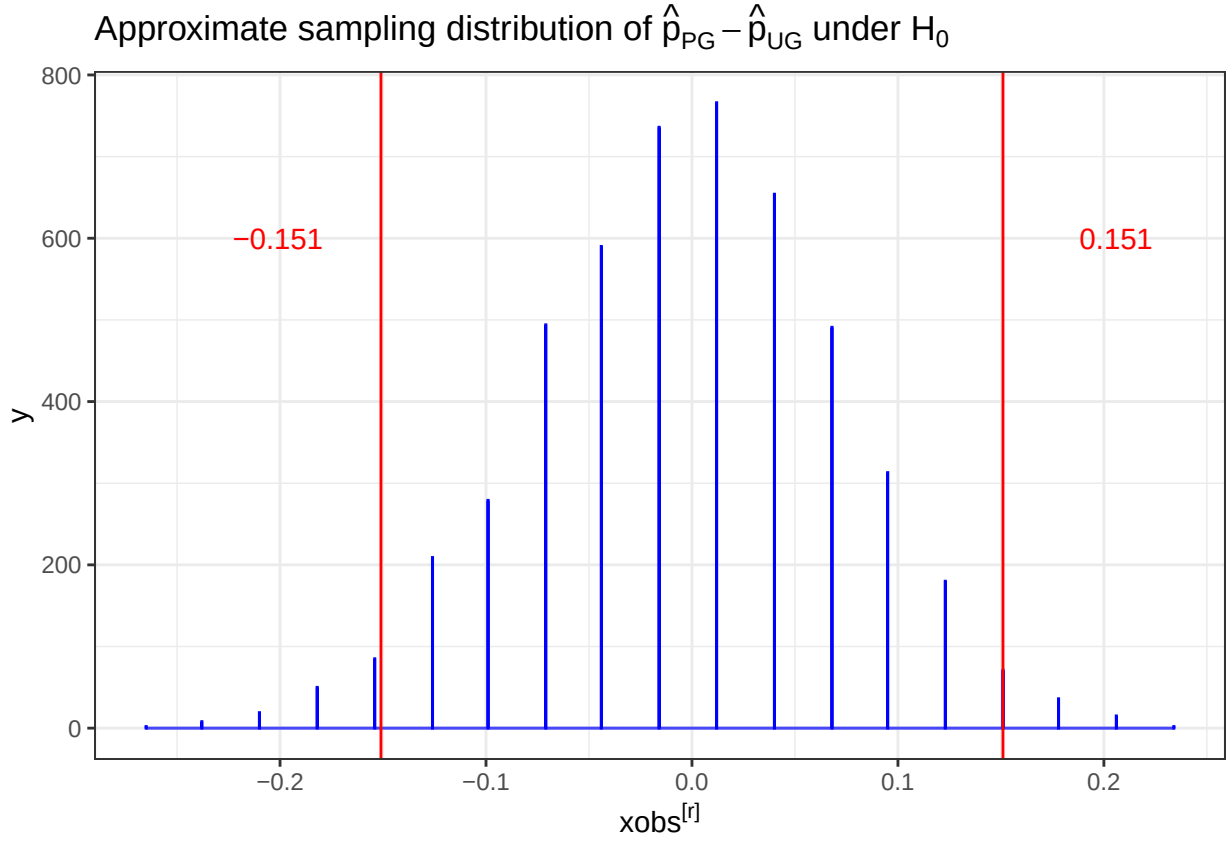
Result	Cohort	Total Students	Proportion
Fail	PG	7	0.146
Fail	UG	43	0.297
Pass	PG	41	0.854
Pass	UG	102	0.703

students successfully passed. The "Proportion" column complements this information by depicting the relative distribution of students within each category as a proportion of the total number of students within their respective cohorts. It is observed that a higher proportion of postgraduate students passed the quiz (i.e. 0.854 compared to 0.703 for undergraduates).

Notably, the table also underscores the importance of providing additional support, with 15% (i.e. 0.146) of postgraduate students and approximately 30% (i.e. 0.297) of undergraduate students facing challenges in meeting the passing criteria.

2.3 Permutation Test & Sampling Distribution

The plot below represents a sampling distribution of proportion differences generated under the assumption that H_0 is true (i.e. that there is no significant difference in pass rates between the two cohorts). The red line shows the proportion difference that we observed from our sample.



2.4 Hypothesis Test

The null hypothesis H_0 can be defined as follows: there is no significant difference in the proportion of students who passed the quiz between the Postgraduate (PG) and Undergraduate (UG) cohorts. Meanwhile, the alternative hypothesis is defined as there is a significant difference in the proportion of students who passed between the "PG" and "UG" cohorts.

This can be written as:

$$H_0 : p_{PG} - p_{UG} = 0 \quad \text{vs.} \quad H_1 : p_{PG} - p_{UG} \neq 0,$$

The p-value from the randomisation test is 0.0576.

Given that this p-value is larger than our chosen significance level of 5% (0.05), it suggests that based on the evidence we have, there is not enough statistical evidence to conclude that there is a significant difference in the proportion of students who passed the quiz between the "PG" and "UG" cohorts. Therefore, we would not reject the null hypothesis $H_0 : p_{PG} - p_{UG} = 0$ as we do not have enough evidence to support the claim of a difference at the 5% significance level.

2.5 Why is the p-value larger than expected?

While it might seem intuitive that such a large difference in proportions should yield a p-value of almost 0, several factors can influence the outcome of the statistical tests, especially in the context of our analysis. Firstly, our sample size of 193 is substantial, but it's important to note that larger samples tend to produce more conservative p-values. In essence, a more conservative p-value is less likely to produce values that are extremely close to 0. This is because with a larger sample size, the statistical test becomes more cautious and demands stronger and more convincing evidence before rejecting the null hypothesis.

Additionally, the presence of data variability may influence the p-value. Our data may exhibit natural variation, making it harder to detect smaller differences as statistically significant. If there is a high degree of variability within the data, even substantial differences in pass rates between postgraduate (PG) and undergraduate (UG) students may not yield p-values close to 0. Moreover, our analysis employs a permutation test, which involves generating a wide range of possible outcomes under the null hypothesis. It's important to recognize that some of these permutations may, by chance, result in relatively large differences between the groups. The p-value considers not only the observed difference but also the distribution of differences under the null hypothesis.

In summary, while the observed difference in pass rates may be substantial and meaningful, the p-value reflects the stringent criteria required to declare a result as statistically significant, considering factors like sample size, data variability, and the specific statistical method employed in our analysis, which aims to assess the significance of pass rate differences between UG and PG cohorts.

3 Bayesian Analysis

3.1 Prior distributions

Since the lecturer has "no real opinion" as to what the distribution of marks for each cohort would be, a Beta prior distribution for the proportion of postgraduate students who passed as well as for the proportion of undergraduate students who passed is most appropriate.

This means that for both cohorts, we would suggest using independent $Beta(\alpha = 1, \beta = 1)$ prior distributions for the two cohorts' pass rates. Reason being, the $Beta(1, 1)$ distribution corresponds to a non-informative prior which essentially assumes that all possible values of the proportion are equally likely. This choice reflects a complete lack of prior knowledge or opinion about the proportion of students who "passed" in each cohort, which matches the lecturer's absence of strong prior opinion/belief about the distribution of marks for each

cohort. Additionally, the Beta distribution is well-suited for modelling proportions and probabilities, since it is bounded between 0 and 1. This aligns with the nature of students' pass rates, as they must fall within this range of 0 to 1.

3.2 Posterior distributions

To calculate the posterior distributions for the proportion of postgraduate students who passed and for the proportion of undergraduate students who passed, we will use the Bayesian framework. Bayes' theorem states that:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

where

- $P(\theta|D)$ is the posterior distribution.
- $P(D|\theta)$ is the likelihood of the data given the parameter.
- $P(\theta)$ is the prior distribution of the parameter.
- $P(D)$ is the marginal likelihood or evidence.

As we've chosen a $Beta(1, 1)$ prior for both cohorts, we can use a binomial likelihood. Since the likelihood is based on the observed data, it follows a binomial distribution for each cohort as the data we have is binary, representing a "pass" or "fail" outcome for each student in each cohort. Additionally, the Beta distribution is the conjugate prior for the binomial likelihood, which means the resulting posterior distribution would also be a Beta distribution. The posterior distributions can be calculated as follows:

For Postgraduate (PG) Students who passed:

- **Likelihood:** $P(\text{Pass}|\theta_{\text{PG}}) \sim \text{Binomial}(n_{\text{PG}}, \theta_{\text{PG}})$, where $n_{\text{PG}} = 41 + 7 = 48$ (total PG students who took the test) and θ_{PG} is the proportion of PG students who passed.
- **Prior:** $P(\theta_{PG}) \sim \text{Beta}(1, 1)$, representing a non-informative prior.
- **Posterior:** $P(\theta_{\text{PG}}|\text{Pass}) \sim \text{Beta}(x + \alpha, n - x + \beta)$
- **Posterior:** $P(\theta_{\text{PG}}|\text{Pass}) \sim \text{Beta}(41 + 1, 48 - 41 + 1)$

Therefore, the posterior distribution for postgraduate students who passed is $P(\theta_{\text{PG}}|\text{Pass}) \sim \text{Beta}(42, 8)$

For Undergraduate (UG) Students who passed:

- **Likelihood:** $P(\text{Pass}|\theta_{\text{UG}}) \sim \text{Binomial}(n_{\text{UG}}, \theta_{\text{UG}})$, where $n_{\text{UG}} = 102 + 43 = 145$ (total UG students who took the test) and θ_{UG} is the proportion of UG students who passed.
- **Prior:** $P(\theta_{\text{UG}}) \sim \text{Beta}(1, 1)$, representing a non-informative prior.
- **Posterior:** $P(\theta_{\text{UG}}|\text{Pass}) \sim \text{Beta}(x + \alpha, n - x + \beta)$
- **Posterior:** $P(\theta_{\text{UG}}|\text{Pass}) \sim \text{Beta}(102 + 1, 145 - 102 + 1)$

Therefore, the posterior distribution for undergraduate students who passed is $P(\theta_{\text{UG}}|\text{Pass}) \sim \text{Beta}(103, 44)$

These posterior distributions represent the updated beliefs about the proportions of students who passed in each cohort based on the observed data and the chosen non-informative $Beta(1, 1)$ prior.

3.3 95% Credible Intervals for PG & UG students who passed

For postgraduate students, there is a 95% chance that the number who passed the quiz is between 35 and 44. As for undergraduate students, there is a 95% chance that the number who passed the quiz is between 91 and 112. In proportionate terms, we observed that there is a 95% chance that the proportion of postgraduate students who passed the quiz is between 0.73 to 0.93. Meanwhile, there is a 95% chance that the proportion of undergraduate students who passed the quiz is between 0.62 to 0.77.

When comparing the two credible intervals, we observe that they do not overlap much, indicating a potential difference in the proportions of postgraduate and undergraduate students who passed the quiz. Additionally, the credible interval for postgraduate students, ranging from 0.73 to 0.93, suggests a higher proportion of 'pass' outcomes among postgraduates with a relatively wide interval. In contrast, the credible interval for undergraduate students, spanning from 0.62 to 0.77, shows a narrower interval with a lower proportion of 'pass' outcomes among undergraduates. These results suggest that postgraduate students may have a higher likelihood of passing the quiz compared to undergraduate students; however, this requires further research and analysis to confirm, as the wider interval for postgraduate students may underscore the greater uncertainty in estimating their 'pass' rate, which could be influenced by factors such as variability and sample size.

3.4 Minimising the posterior expected squared error loss and the corresponding credibility factors

The estimator that minimises the posterior expected squared error loss for the proportion of students who passed for a given cohort is the posterior mean $\hat{\theta}_{\text{Bayes}} = E(\theta|X)$. For the undergraduate cohort, the posterior mean for the proportion of students who passed is approximately 0.701, implying that we estimate approximately 70.1% of undergraduate students in the cohort passed the quiz. Meanwhile, the posterior mean for proportion of postgraduate students who passed is approximately 0.84, suggesting that approximately 84% of postgraduate students is estimated to pass. These posterior means are in a form that linearly combines a purely data-based estimator and some prior quantity, with the data-based estimator being the sample mean of each cohort and the prior quantity being the prior distribution previously chosen for the proportion of students who passed in each cohort.

The credibility factor for the undergraduate cohort is 0.986, inferring that approximately 98.6% of the posterior estimate is contributed by the data-based estimator, after observing the data for the undergraduate cohort and combining it with the prior information. Meanwhile, the credibility factor for the postgraduate cohort is 0.96, suggesting that approximately 96% of the posterior estimate is contributed by the data-based estimator. Therefore, we can conclude that the data-based estimator (sample mean) contributes significantly to the posterior estimate for both cohorts.

3.5 Minimising the posterior expected absolute error loss

```
postug_median <- qbeta(0.5, 103, 44)  # undergraduate
```

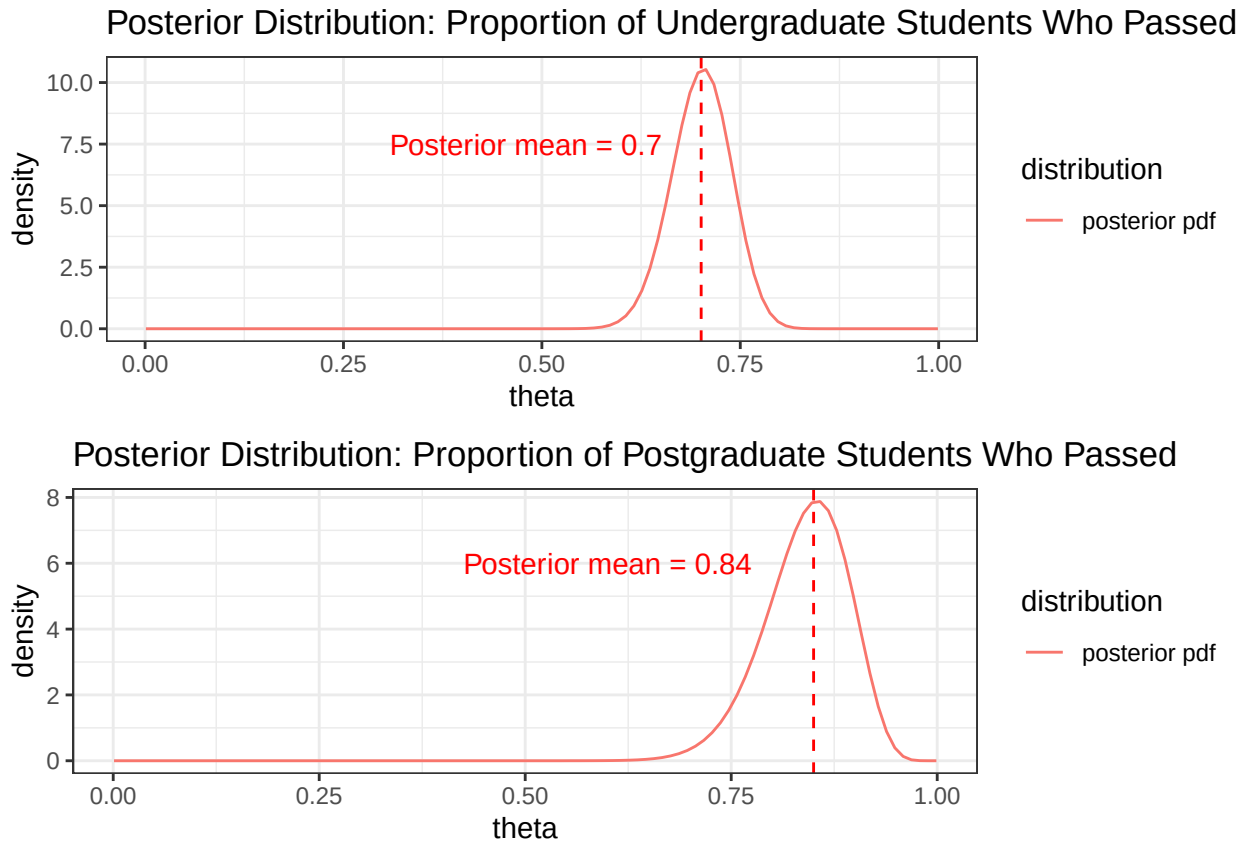
```
postpg_median <- qbeta(0.5, 42, 8)  # postgraduate
```

The estimator that minimises the posterior expected absolute error loss for the proportion of students who passed for a given cohort is the posterior median. For the undergraduate cohort, the posterior median is calculated to be 0.702. This suggests that based on our Bayesian analysis and the available data and prior information, we estimate that the proportion of undergraduate students who passed the quiz is approximately 0.7 with a 50% probability. Meanwhile, the posterior median for the postgraduate cohort is 0.845. This infers that there is a 50% probability of the proportion of postgraduate students who passed being approximately 0.85.

3.6 Visualising Posterior Distribution: Students who "Passed"

Based on the probability density functions (pdf) of posterior distributions for the proportion of undergraduate and postgraduate students who passed, we observe that the posterior distribution for the proportion of undergraduate students who passed appears to be roughly symmetric and unimodal. The peak of the distribution, which aligns with the posterior mean (0.701 or 70%), represents the most likely proportion of undergraduates who passed. This symmetric shape suggests that the uncertainty in the estimate is relatively balanced on both sides of the peak.

Meanwhile for postgraduate students, the posterior distribution for the proportion of postgraduate students who passed appears to be left-skewed. The mean posterior distribution (i.e. 0.85) also represents the most likely proportion of postgraduates who passed. However, this left skewness indicates that while 0.85 (85%) is the most probable estimate for the proportion of postgraduates who passed, there is a possibility of lower values being the estimate eventhough they are less likely. This therefore suggests that there is a certain degree of uncertainty/variability in the estimate, as it's also possible that the true proportion of postgraduate students who passed could be lower than the 0.85 estimate.



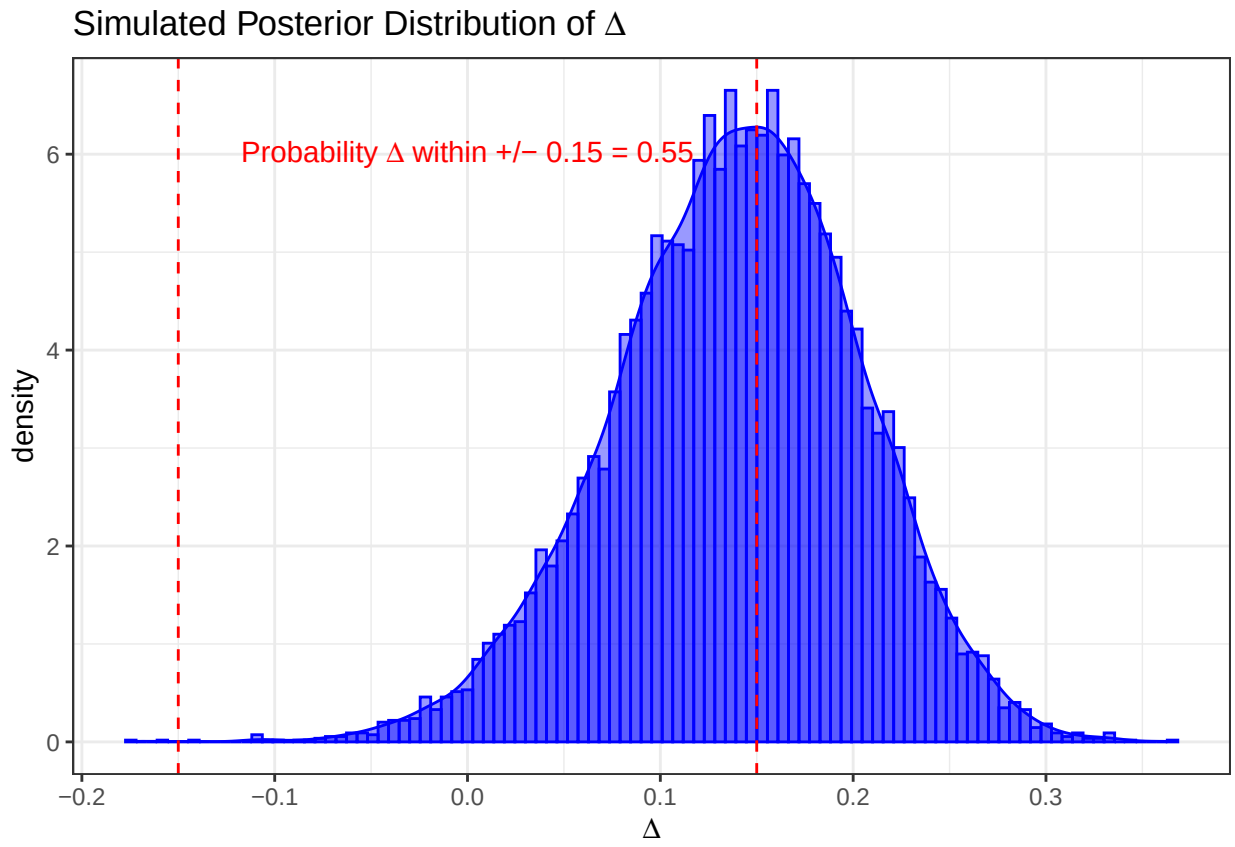
3.7 Recommending Priors for next semester

We would advise the lecturer to use the posterior α and β parameters from the current semester (i.e. $P(\theta_{\text{UG}}|\text{Pass}) \sim \text{Beta}(103, 44)$ and $P(\theta_{\text{PG}}|\text{Pass}) \sim \text{Beta}(42, 8)$) as the new priors for the next semester's analysis. This is because by using the current semester's posterior parameters as the new priors, the lecturer would be incorporating the information and insights gained from the current semester's data into the next semester's analysis, which reflects the Bayesian principle of updating beliefs based on new evidence. Additionally, by using the posterior parameters as priors, the lecturer is acknowledging the uncertainty in the

parameter estimates since the posterior distribution represents a summary of the current semester's data and the lecturer's prior beliefs.

3.8 Visualising Simulated Posterior Distribution: Postgrad vs. Undergrad Pass Rate

Based on the simulated posterior distribution of the difference in proportion of postgraduate students who passed and the proportion of undergraduate students who passed (Δ), we can observe that the posterior mean of (Δ) is approximately 0.139. This means that on average, we can expect that the percentage of postgraduate students who passed is estimated to be about 13.9% (approximately 14%) higher than the percentage of undergraduate students who passed, as indicated by the positive value of Δ.



3.9 Probability of Grade Difference is within 0.15

According to our calculations, the probability that the difference in the grades of the two cohorts is within +/- 0.15 is 0.55 (i.e. 55%). This suggests that approximately 55% of the simulated differences in grades between the two cohorts fall within the range of -0.15 to 0.15. Based on this, we can infer that there is a relatively high probability (i.e. 55%) that the difference in grades between the two cohorts is within +/- 0.15. This also suggests that the two cohorts have almost similar grade distributions, since a substantial portion of the grade differences falls within this narrow range.