08_A1_Task3_Beta_Beta

Group 08

2023-09-18

Questions about Task 3

- For the prior, do we just make it up and justify, or is there a concrete way to do this??

Task 1 - Fit a Distribution

Tidying the Data

Task 2 - Are Post Grad Students better?

New Variable to Data - Pass or Fail

Relevant Proportion Table

Task 3 - Bayesian Analysis

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 the Undergraduates that Passed

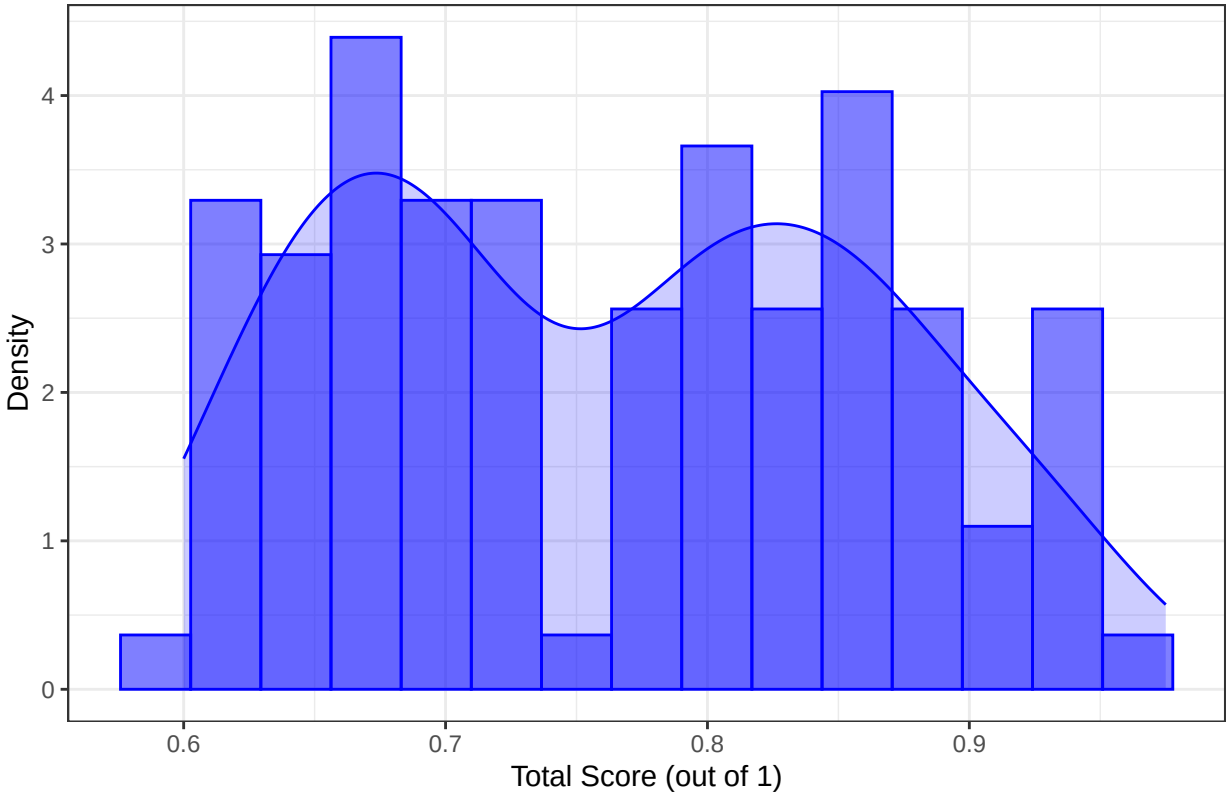


Figure 9: Distribution of the Post graduate that Passed

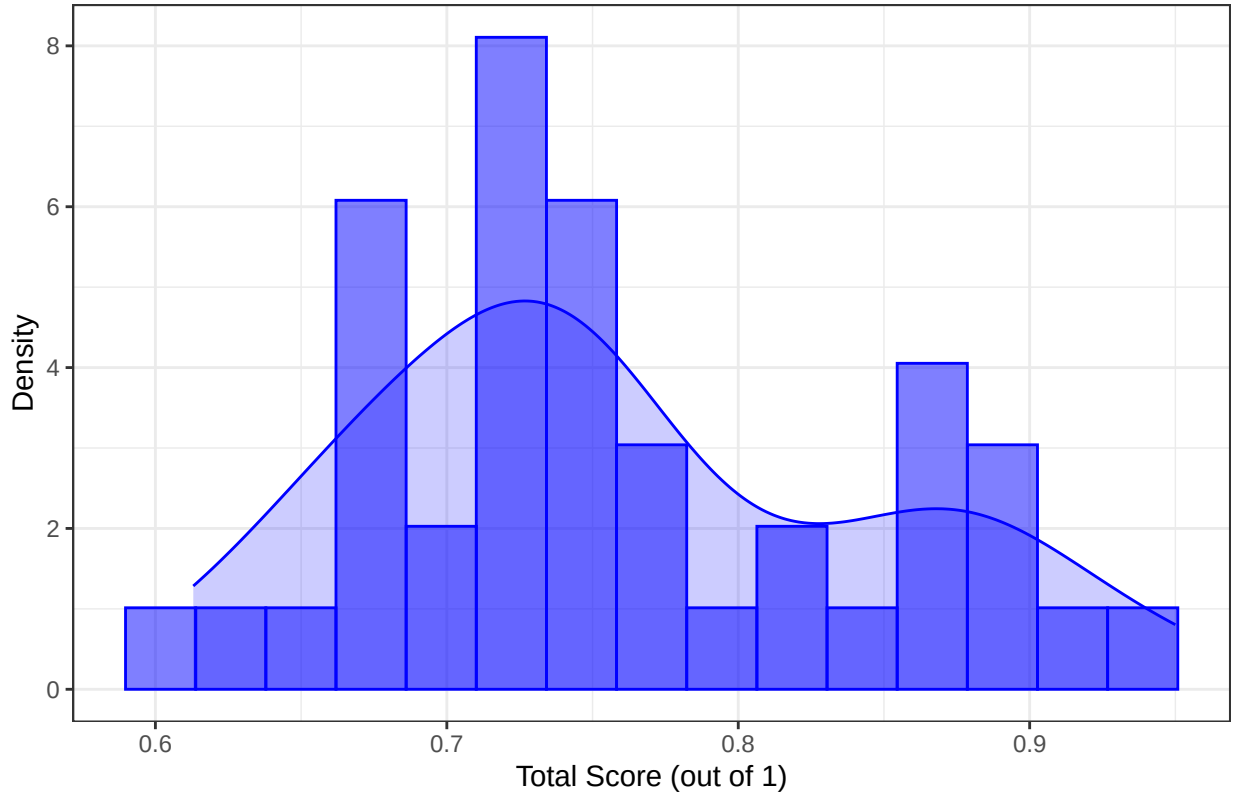


Figure 8 illustrates the distribution for undergraduate students that "passed" the quiz. As can be seen, the distribution seems to be a continuous bi-modal distribution. To an extent, it looks like similar to a continuous uniform distribution and as such, an appropriate prior distribution for the proportion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

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 distributon 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 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 post graduate students that passed lies within the interval [0.7275785,0.9267772], based on the data and our prior beliefs.

There is a 95% probability that the true value of the mean of the post graduate students that passed lies within the interval [0.6244538,0.7717348], based on the data and our prior beliefs.

As can be seen, for the post graduate students, the interval is larger than that for the undergraduate students. This indicates greater disparity and more uncertainty/variability with the post graduate students mean, than the undergraduate students mean.

Q5: 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 ($\hat{\mu}_{UG}$) = 0.7006803

Credibility Factor (Z) = 0.9863946

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 0.7006803 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:

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

Q6: 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,

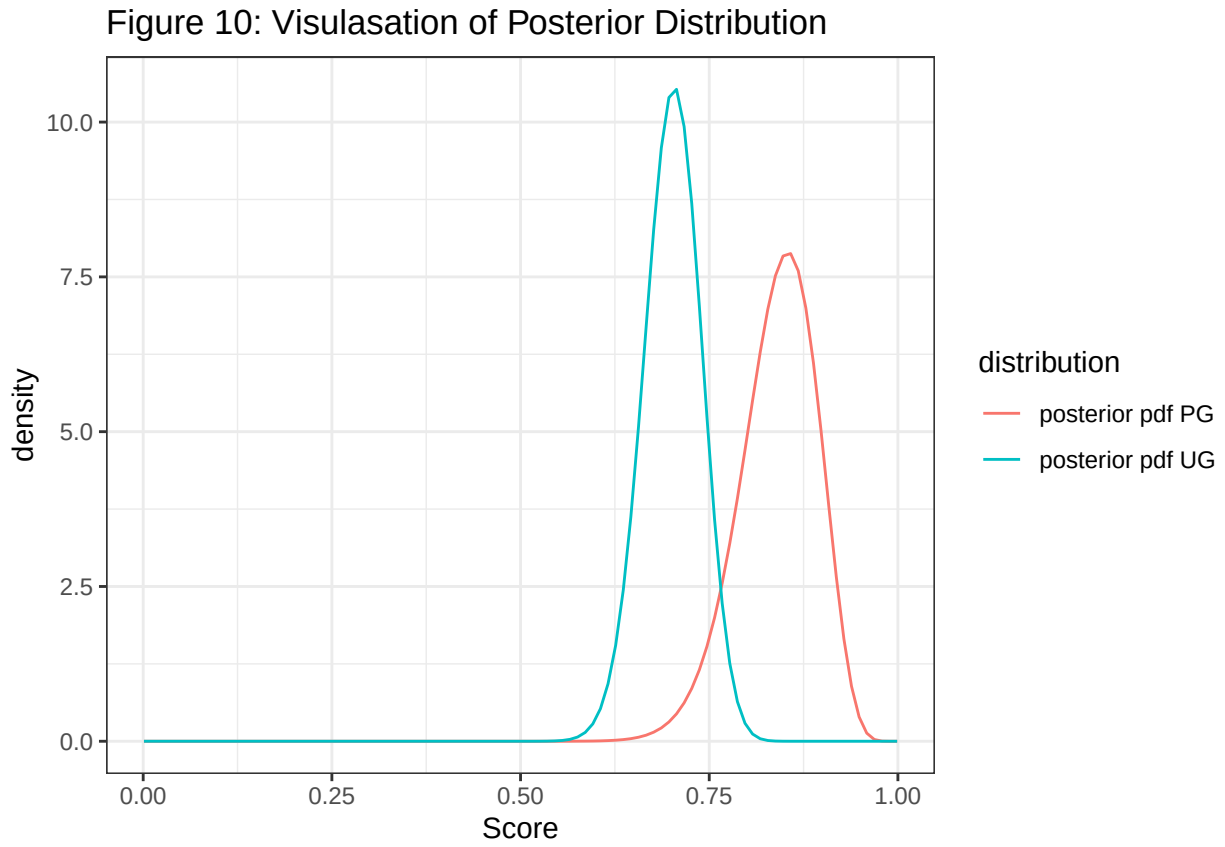
The estimate for for the undergraduate cohort: 0.7015926

The estimate for for the post graduate cohort: 0.8445491

The estimates that minimize the posterior expected absolute error loss are 0.7015926 and 0.8445491 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:

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

Q7: 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.

Q8: 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 $X \sim beta(\alpha, \beta)$.

Specifically,

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

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

Q9: The posterior distribution of the difference in the proportion [4 marks]

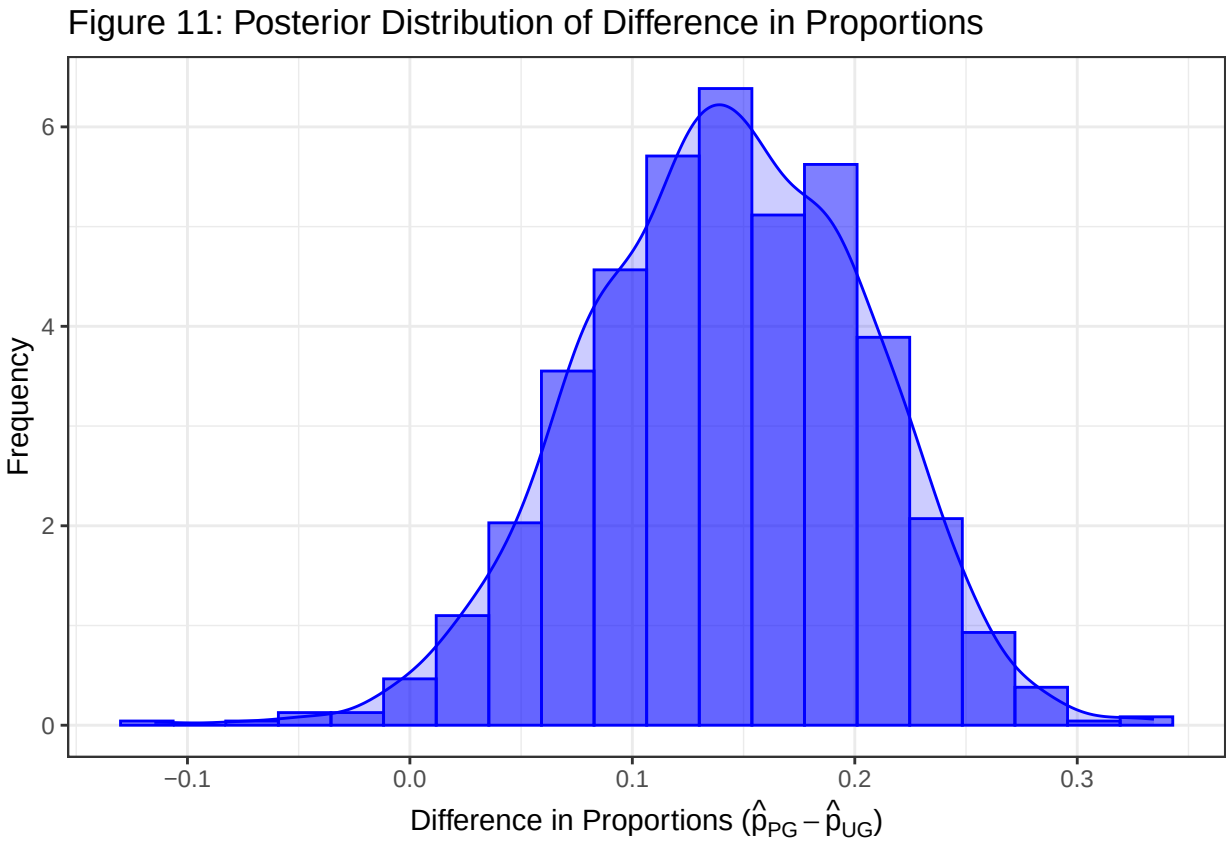
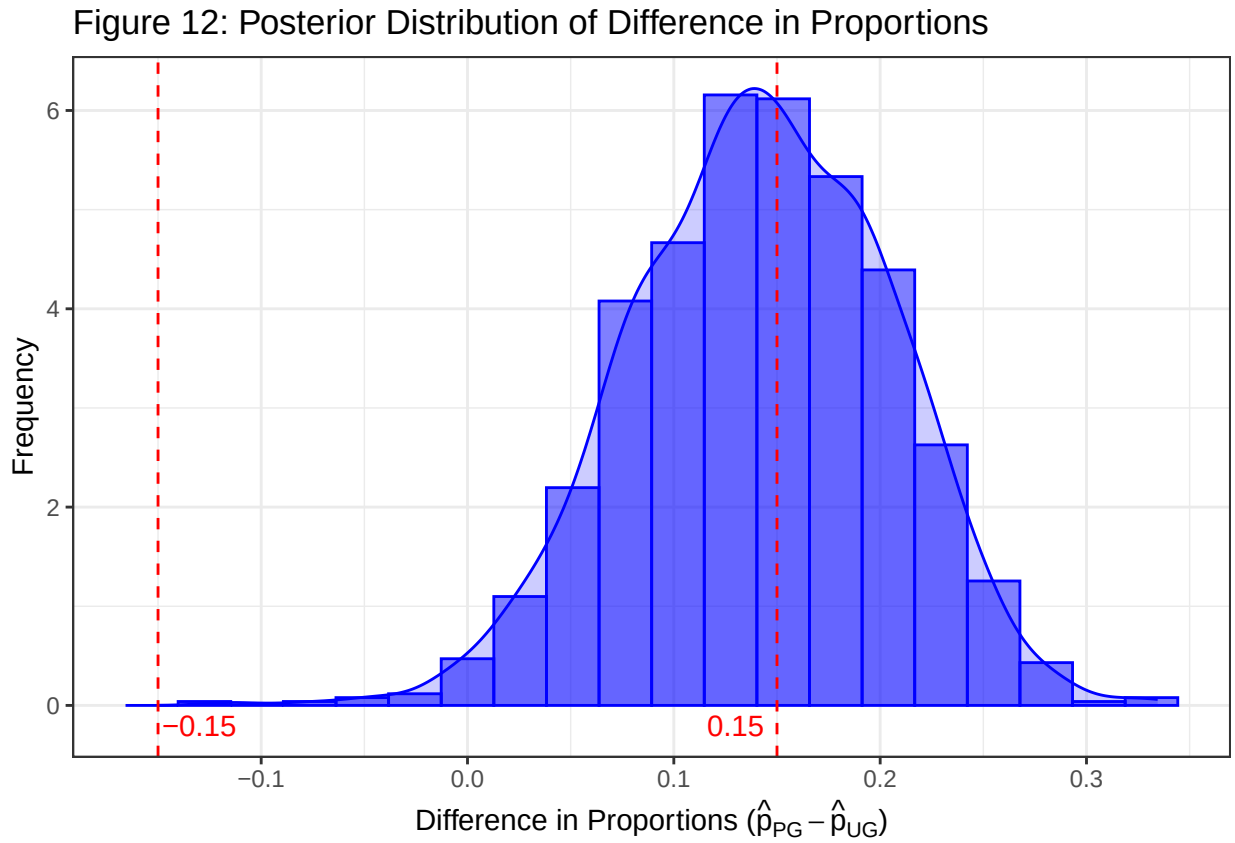


Figure 11 shows that the posterior distribution of the difference in proportions is somewhat normally distributed (slightly negatively skewed). This distribution is not around 0.0, indicating that the average proportion of postgraduate students that passed is greater than the proportion of undergraduate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

Q10: 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 0.549. 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.
