08_A1_Task3

Group 08 - Jason Abi Chebli (31444059), Sovathanak Measn (), Neev Bhandari (), Farrel Wiharso ()

2023-09-18

```
library(formatR)
library(tidyverse)
library(kableExtra)
library(broom)
library(MASS)

GradesData <- read.csv("GradesData.csv")
```

Questions about Task 3

- For the prior, do we just make it up and justify, or is there a concrete way to do this??
- What Alpha and beta do we use
- How do we minimises the posterior expected squared error loss? Is this week 9 stuff?
- When calculating the difference, what is the x and n value ?? do we sum... I subtracted

Task 1 - Fit a Distribution

Tidying the Data

```
# Modify the data to ensure that only genuine (non zero)
# attempts are analysed.
TidyGradesData <- GradesData |>
    filter(Total != 0)
```

Task 2 - Are Post Grad Students better?

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

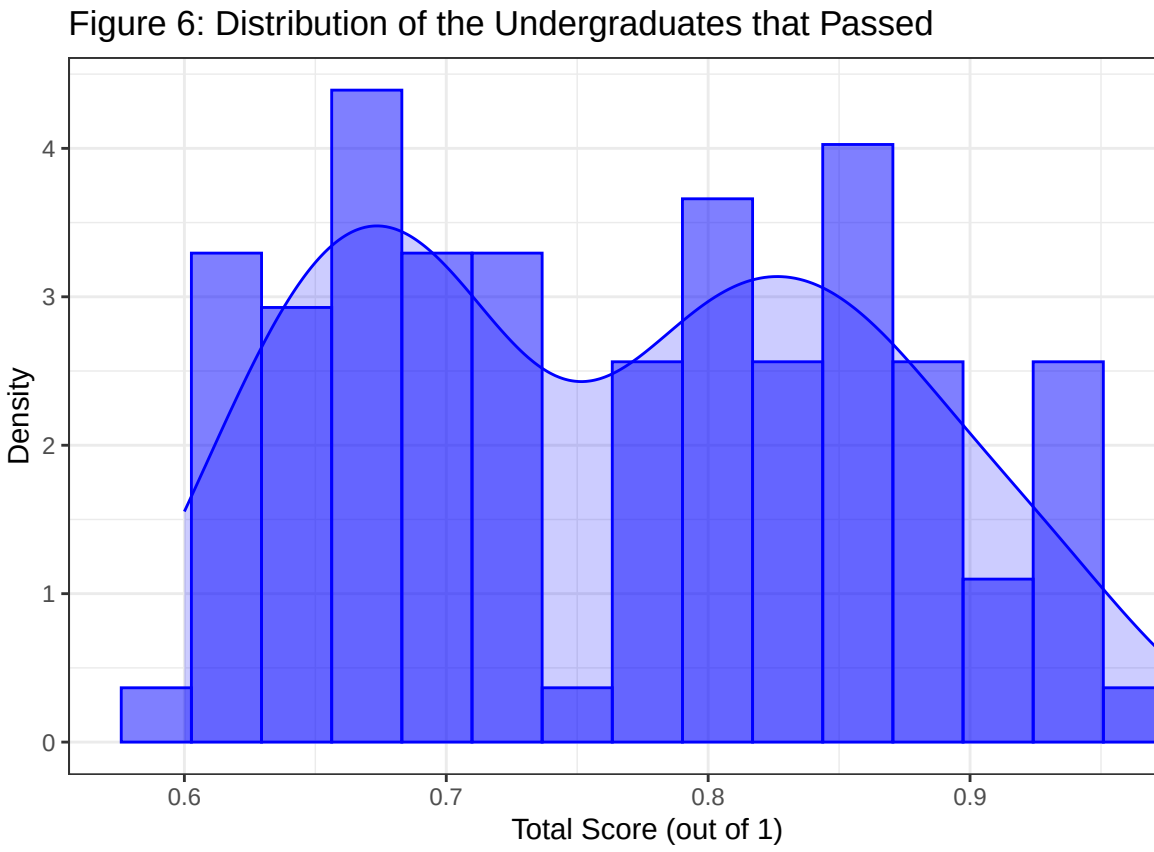
New Variable to Data - Pass or Fail

```r
numResults <- TidyGradesDataWithResults |>
    group_by(Cohort, Result) |>
    summarize(Sum = n()) |>
    pivot_wider(id_cols = Cohort, names_from = Result, values_from = Sum) |>
    mutate(TotalSum = FAIL + PASS)
propResults <- numResults |>
    group_by(Cohort) |>
    mutate(prop_fail = FAIL/TotalSum, prop_pass = PASS/TotalSum) |>
    summarise(Cohort, prop_fail, prop_pass)
```

Relevant Proportion Table

Task 3 - Bayesian Analysis

```r

TidyGradesDataWithResults |>
    filter(Result == "PASS", Cohort == "UG") |>
    ggplot(aes(x = Total, y = after_stat(density))) + geom_histogram(colour = "blue",
    fill = "blue", alpha = 0.5, bins = 15) + geom_density(colour = "blue",
    fill = "blue", alpha = 0.2) + labs(x = "Total Score (out of 1)",
    y = "Density") + ggtitle("Figure 6: Distribution of the Undergraduates that Passed") +
    theme_bw()
```



Prior Distributions

```
TidyGradesDataWithResults |>
    filter(Result == "PASS", Cohort == "PG") |>
    ggplot(aes(x = Total, y = after_stat(density))) + geom_histogram(colour = "blue",
    fill = "blue", alpha = 0.5, bins = 15) + geom_density(colour = "blue",
    fill = "blue", alpha = 0.2) + labs(x = "Total Score (out of 1)",
    y = "Density") + ggtitle("Figure 7: Distribution of the Undergraduates that Passed") +
    theme_bw()
```

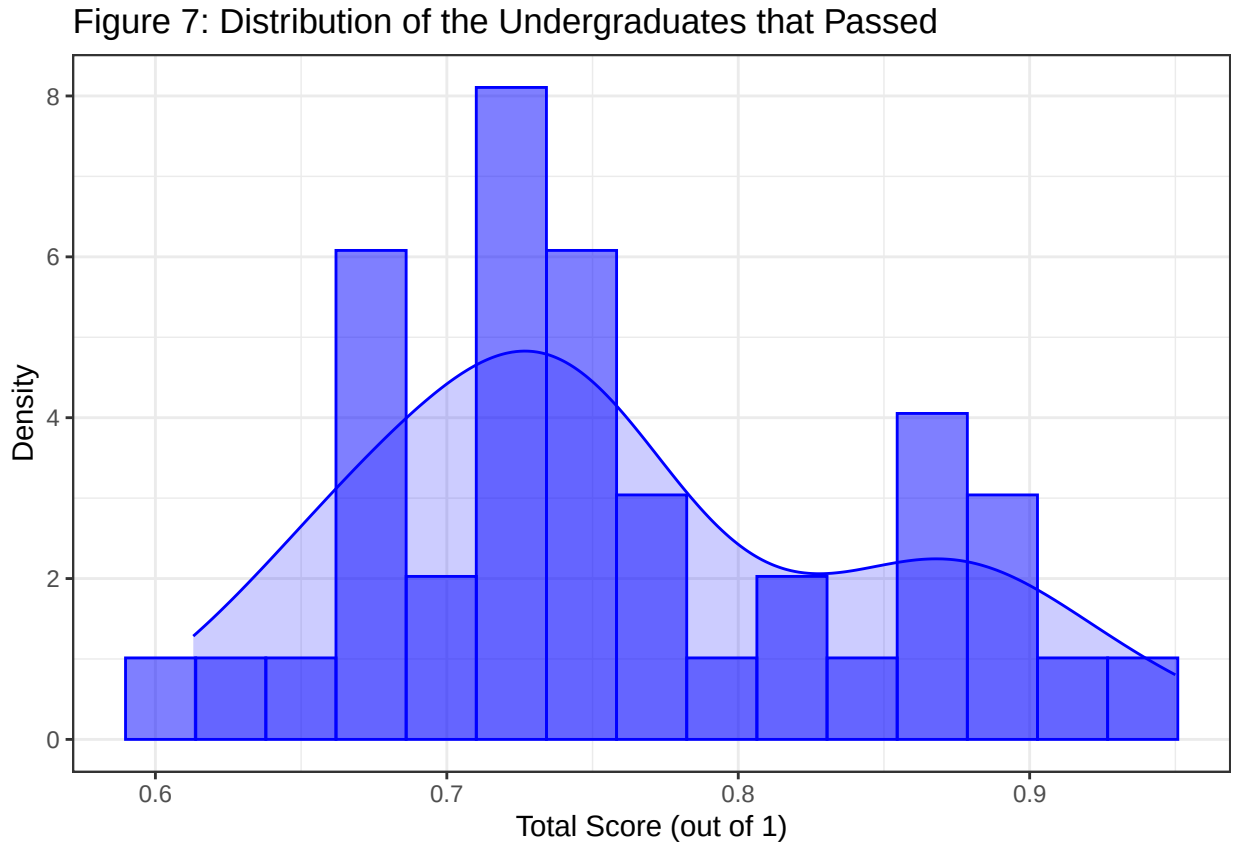


Figure 6 illustrates the distribution for undergraduate students that "passed" the quiz. As can be seen, it is a continuous bi-modal distribution. To an extent, it looks like similar to a continuous uniform distribution and as such, an appropriate prior distribution for the proportion of undergraduate students who "passed" could be a beta distribution.

(Comment: maybe it is normal????)

Figure 7, 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 and approaching a uni-modal distribution. As such, an appropriate prior distribution for the proportion of postgraduate students who "passed" could be a gamma distribution.

Posterior Distributions 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 who "passed", given that the prior distribution is assumed to have 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.

Now, for the proportion of postgraduate students who "passed", given that the prior distribution is assumed to have a gamma distribution, this means that the posterior can either be Poisson or exponential. Similarly, Poisson is a discrete distribution and Exponential is a continuous distribution. Hence, given that the proportion is a continuous distribution, the posterior distribution for the proportion of postgraduate students who "passed" must be a exponential distribution.

In Summary: UG: Beta-Binomial ; PG: Gamma-Exponential

```
beta_binomial <- function(n, x, alpha = 1, beta = 1) {
    atil <- alpha + x
    btil <- beta + n - x
    z <- n/(alpha + beta + n)
    out <- list(alpha_tilde = atil, beta_tilde = btil, credibility_factor = z)
    return(out)
}
```

```
gamma_exponential <- function(n, x, alpha = 1, beta = 1) {
    atil <- alpha + n
    btil <- beta + n * x
    xbar <- mean(x)
    z <- (n * xbar)/(n * xbar + beta)
    out <- list(alpha_tilde = atil, beta_tilde = btil, credibility_factor = z)
    return(out)
}
```

```
alpha <- 1
beta <- 1

n_PG <- numResults$TotalSum[numResults$Cohort == "PG"]
x_PG <- numResults$PASS[numResults$Cohort == "PG"]

n_UG <- numResults$TotalSum[numResults$Cohort == "UG"]
x_UG <- numResults$PASS[numResults$Cohort == "UG"]

out_PG <- gamma_exponential(n_PG, x_PG, alpha, beta)
out_UG <- beta_binomial(n_UG, x_UG, alpha, beta)

credible_interval_PG <- qgamma(c(0.025, 0.975), shape = out_PG$alpha_tilde,
    rate = out_PG$beta_tilde)
credible_interval_UG <- qbeta(c(0.025, 0.975), shape1 = out_UG$alpha_tilde,
    shape2 = out_UG$beta_tilde)
```

95% Credibility Intervals We are 95% confident that the true value of the parameter for the post graduate students lies within the interval [0.0184106,0.0323215], based on the data and our prior beliefs.

We are 95% confident that the true value of the parameter for the undergraduate students lies within the interval [0.6244538,0.7717348], based on the data and our prior beliefs.

Really confused :((

```
# Determine the prior mean
prior_mean_UG <- alpha/(alpha + beta)  # Beta distribution
prior_mean_PG <- alpha/beta  # Gamma distribution

# Extract the credibility factor
credibility_factor_UG <- out_UG$credibility_factor
credibility_factor_PG <- out_PG$credibility_factor

# Yes, it is this in a form that linearly combines a purely
# data-based estimator and some prior quantity

bayesian_estimate_UG <- credibility_factor_UG * (x_UG/n_UG) +
    (1 - credibility_factor_UG) * prior_mean_UG
bayesian_estimate_PG <- credibility_factor_PG * (1/(mean(x_PG))) +
    (1 - credibility_factor_PG) * (alpha/beta)
```

The estimator that minimises the posterior expected squared error loss 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:

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

Meanwhile, the form of the estimator for a gamma-exponential is:

$$\hat{\mu}_{PG} = Z \times (1/\bar{x}) + (1 - Z) \times (\alpha/\beta)$$

For the undergraduate cohort:

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

Credibility Factor (Z) = 0.9863946

For the post graduate cohort:

Estimate ($\hat{\mu}_{PG}$) = 0.0248857

Credibility Factor (Z) = 0.9994921

```
# For UG cohort (Beta-Binomial)
posterior_samples_UG <- rbeta(n_UG, out_UG$alpha_tilde, out_UG$beta_tilde)
```

```r
# For PG cohort (Gamma-Exponential)
posterior_samples_PG <- rexp(n_PG, rate = out_PG$beta_tilde)

# Calculate the posterior median for UG
posterior_median_UG <- quantile(posterior_samples_UG, 0.5)

# Calculate the posterior median for PG
posterior_median_PG <- quantile(posterior_samples_PG, 0.5)
```

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

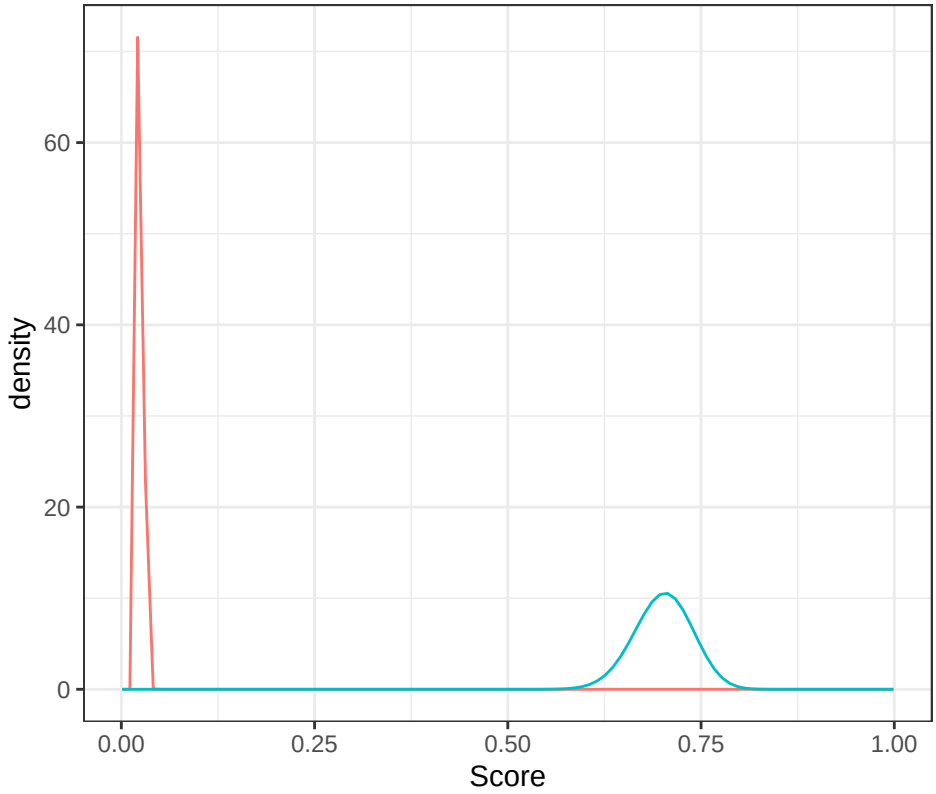
Hence,

The estimate for for the undergraduate cohort: 0.7072582

The estimate for for the post graduate cohort: 2.9056151×10^{-4}

```r
cbbPal <- c(black = "#000000", orange = "#E69F00", ltblue = "#56B4E9",
    "#009E73", green = "#009E73", yellow = "#F0E442", blue = "#0072B2",
    red = "#D55E00", pink = "#CC79A7")
cbbP <- cbbPal[c("orange", "blue", "pink")]  #choose colours for p

x_values <- seq(0.001, 0.999, length.out = 100)
postUG <- dbeta(x = x_values, out_UG$alpha_tilde, out_UG$beta_tilde)
postPG <- dgamma(x = x_values, out_PG$alpha_tilde, out_PG$beta_tilde)

df <- tibble(Score = x_values, `posterior pdf UG` = postUG, `posterior pdf PG` = postPG)
df_longer <- df %>%
    pivot_longer(-Score, names_to = "distribution", values_to = "density")
p1 <- df_longer %>%
    ggplot(aes(x = Score, y = density, colour = distribution,
        fill = distribution)) + geom_line() + scale_fill_manual(values = cbbP) +
    theme_bw() + labs(title = "Figure 8: Visulasation of Posterior Distribution")
p1
```

Figure 8: Visulasation of Posterior Distribution



Visualise the posterior distribution

Which prior to use in future

Even though our prior should 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 posterior distirbution, i.e., the beta distribution for the undergraduate cohort and the gamma distribution for the postgraduate cohort. This is because we now are updating our "beliefs".

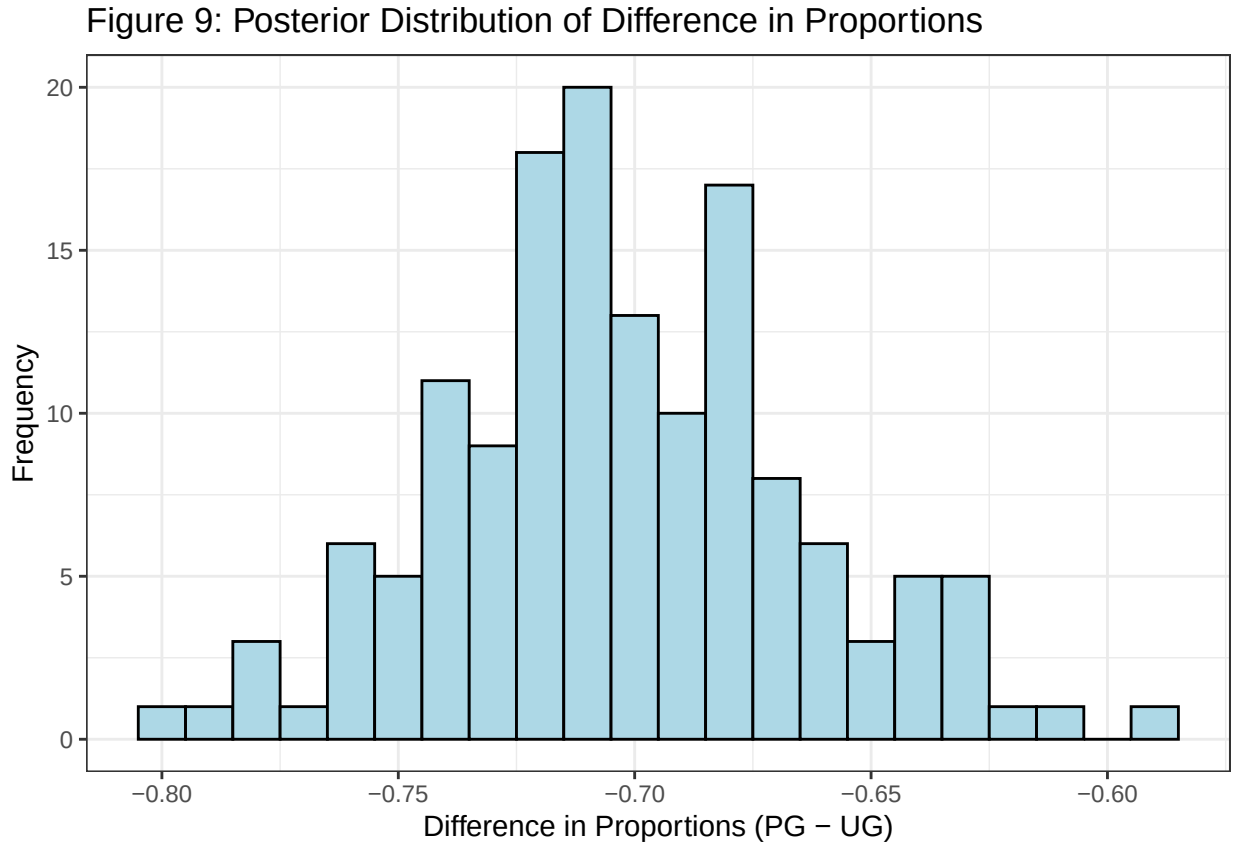
```
alpha <- 1
beta <- 1

n_PG <- numResults$TotalSum[numResults$Cohort == "PG"]
x_PG <- numResults$PASS[numResults$Cohort == "PG"]

n_UG <- numResults$TotalSum[numResults$Cohort == "UG"]
x_UG <- numResults$PASS[numResults$Cohort == "UG"]

n_diff <- n_PG - n_UG
x_diff <- x_PG - x_UG

out_diff <- beta_binomial(n_diff, x_diff, alpha, beta)

credible_interval_diff <- qbeta(c(0.025, 0.975), out_diff$alpha_tilde,
    out_diff$beta_tilde)
```

```
credible_interval_diff
```

The posterior distribution of the difference in the proportion

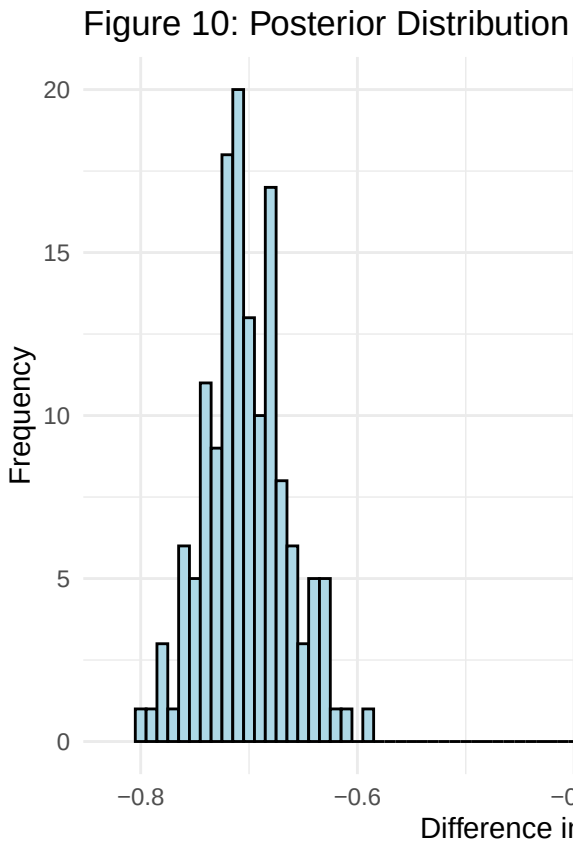
```
## [1] NaN NaN
```

```
posterior_diff_samples <- posterior_samples_PG - posterior_samples_UG

# Create a data frame with posterior samples
posterior_data <- data.frame(Difference = posterior_diff_samples)

# Visualize the posterior distribution
posterior_data %>%
    ggplot(aes(x = Difference)) + geom_histogram(binwidth = 0.01,
    fill = "lightblue", color = "black") + labs(title = "Figure 9: Posterior Distribution of Difference
    x = "Difference in Proportions (PG - UG)", y = "Frequency") +
    theme_bw()
```




```
probability_within_range <- posterior_data %>%
    summarize(Probability = mean(abs(Difference) <= 0.15))
```

```
posterior_data %>%
    ggplot(aes(x = Difference)) + geom_histogram(binwidth = 0.01,
    fill = "lightblue", color = "black") + geom_vline(xintercept = c(-0.15,
    0.15), linetype = "dashed", color = "red") + labs(title = "Figure 10: Posterior Distribution of Dif
    x = "Difference in Proportions (PG - UG)", y = "Frequency") +
    theme_minimal()
```



The probability that the difference in the grades of the 2 cohorts

The probability that the difference in the grades of the 2 cohorts is within 0.15 is 0. As can be seen in Figure 10,