08_A1_Task2

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 2

- Do we still use the tidy data (197 observations) for Task 2 and 3??
- We don't need the "select" function???
- for the proportion, do you want a table just of proportion, or with the other values as well?
- how many digits to display? 3 digits? up to us?
- how do we know when we have done enough bins ??
- One sided or two sided test?? in class, we did a two sided with CBD even though we knew that it should be a one sided?

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)

propResults |>
    rename(`Fail Proportion` = prop_fail, `Pass Proportion` = prop_pass) |>
    kable(caption = "Table 1: Proportion of Fail and Pass Results by Cohort",
        digits = 3) |>
    kable_styling(latex_options = "hold_position")
```

Table 1: Table 1: Proportion of Fail and Pass Results by Cohort

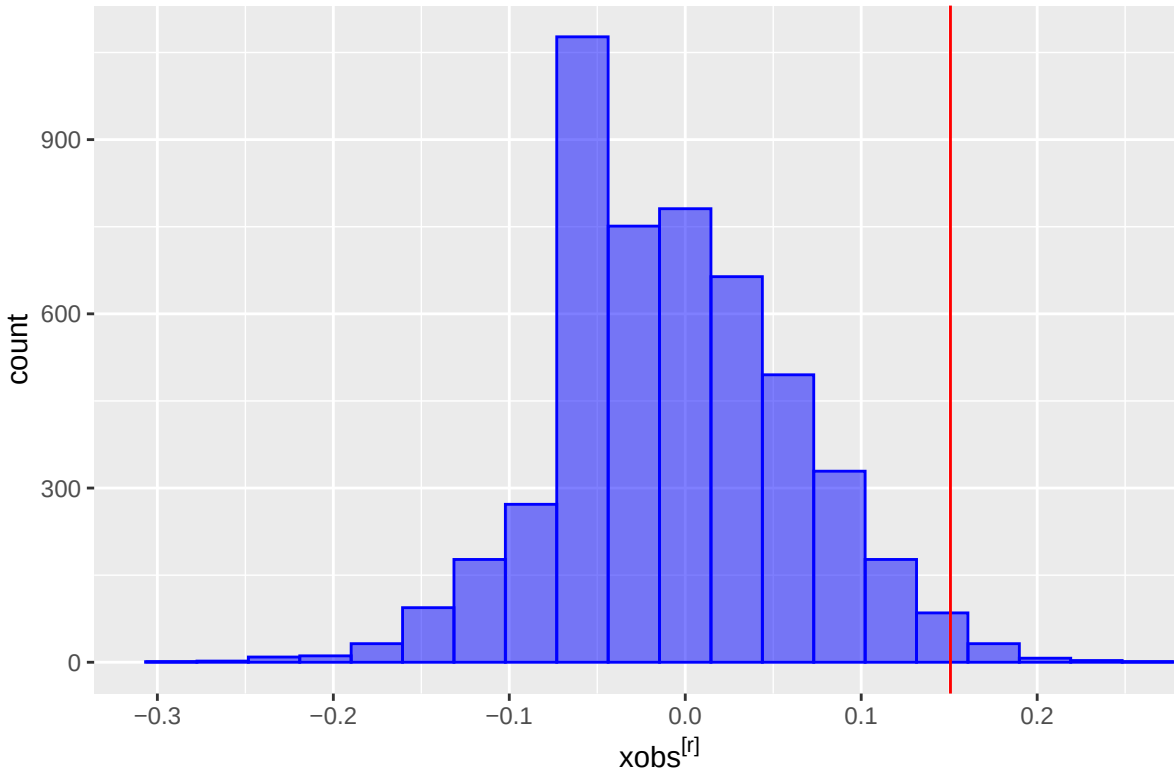
Cohort	Fail Proportion	Pass Proportion
PG	0.146	0.854
UG	0.297	0.703

Relevant Proportion Table 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.

```r
xobs <- propResults$prop_pass[propResults$Cohort == "PG"] - propResults$prop_pass[propResults$Cohort ==
    "UG"]


n <- nrow(TidyGradesDataWithResults)

# set the number of replications for test
R <- 5000

# set up variable array (a vector with R spaces all filled
# with NA) to be used to store the replicated 'xobs' values
# after shuffling
Rxobs <- array(dim = R)
# <left blank for part 7b>
set.seed(2023)

# initialise randomisation sample with the original data
# will replace inside loop with each permutation
RTidyGradesDataWithResults <- TidyGradesDataWithResults

# Use for-loop to generate R replications of the test
```

```r
# statistic Note the format. All commands inside the curly
# brackets occur for each value of r in the sequence 1:R

for (r in 1:R) {

    # for iteration r shuffle the gender variable in the
    # data file
    TidyGradesDataWithResults <- TidyGradesDataWithResults |>
        mutate(Result = sample(TidyGradesDataWithResults$Result,
            n, replace = FALSE))

    # calculate the summary table GDSS3 for the shuffled
    # tibble
    RPropResults <- TidyGradesDataWithResults %>%
        group_by(Cohort, Result) %>%
        tally() %>%
        ungroup() %>%
        pivot_wider(names_from = Result, values_from = n) %>%
        mutate(TotalSum = FAIL + PASS) %>%
        mutate(prop_fail = round(FAIL/TotalSum, digits = 3),
            prop_pass = round(PASS/TotalSum, digits = 3))

    # save the 'xobs' for the shuffled data in the rth
    # element of the Rxobs vector (array)
    Rxobs[r] <- RPropResults$prop_pass[propResults$Cohort ==
        "PG"] - RPropResults$prop_pass[propResults$Cohort ==
        "UG"]

    # close the for-loop with a curly bracket.
}

# Proportion of Rxobs at least as big as xobs - approximate
# p-value
pval <- sum(Rxobs >= xobs)/R
options(digits = 4)
# use this if you wish to have the p-value printed under
# the code chunk I have commented it out as we will use it
# inline laterf
```

Figure 5: Approximate sampling distribution of $\hat{p}_{PG} - \hat{p}_{UG}$ under H_0



Permutation Test

What is the result of this test? To determine whether post graduate students are better, we want to see if the proportion of post graduate 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 5. The plot in Figure 5 represents a sampling distribution of proportion differences generated under the assumption that H0 is true (i.e.that post graduate 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 0.0256. This indicates that at a 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 5% and 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.

Why p-value is not 0? Such a large difference in proportions does not have a p-value of almost 0. 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 discrepency the smaller the p-value would be.
2. The larger the sample size, in general holding all else constant, the smaller the p-value. Our sample size is 193. If this increased a lot more, the p-value would most likely approach 0 - assuming all else held constant.
3. ???
4. ???