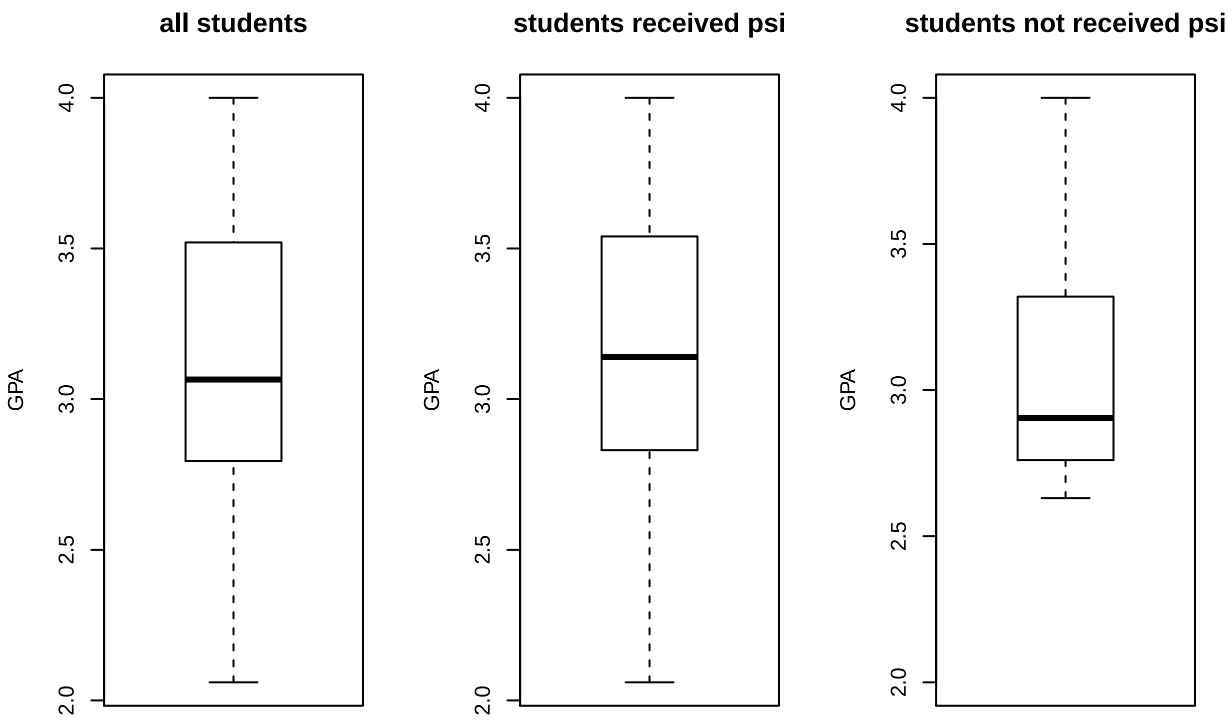
Assignment 3 Exercise 2

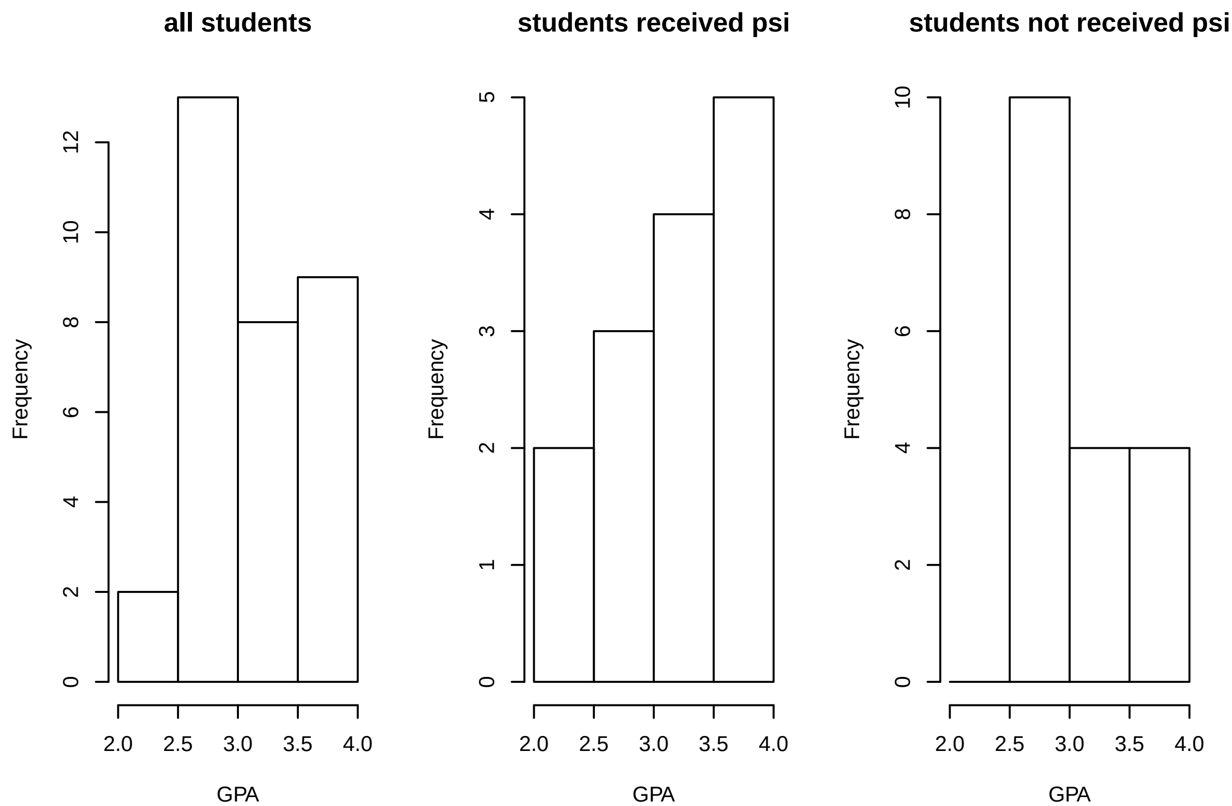
Sophie

3/15/2020

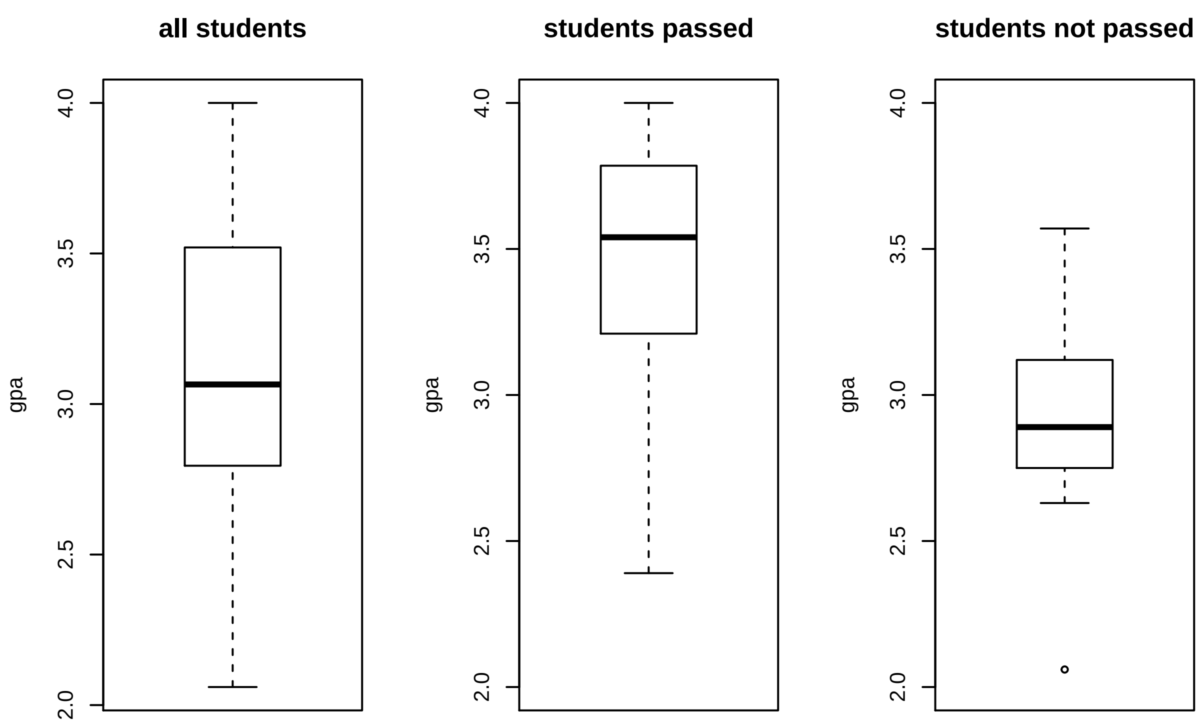
Exercise 2

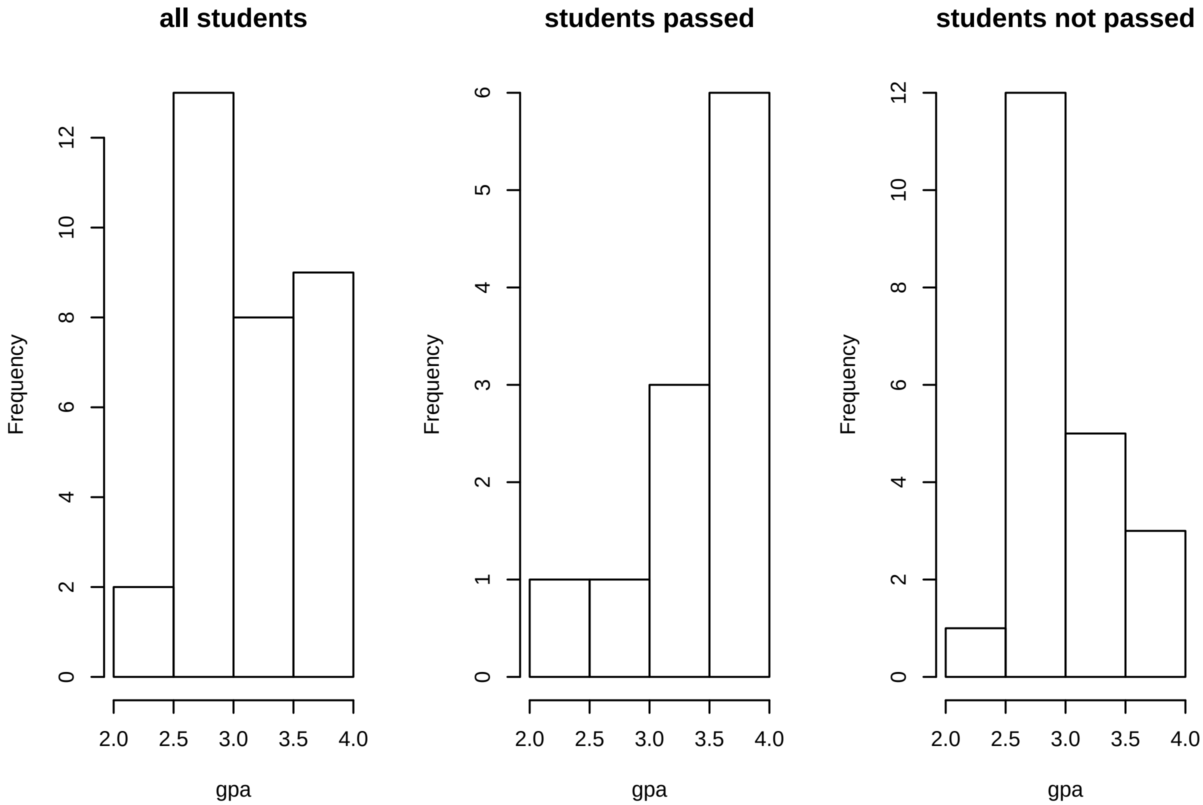
a) We study the data by exploring all combinations of the variables. First, we investigate the relation between the variables psi and gpa. We are interested whether the students that receive psi have a similar GPA to the students not receiving psi. We visualized the data in the boxplots below. We observe that the GPAs of all students is evenly distributed. The same applies to the GPAs of the students who received psi, however, the boxplot is positioned slightly higher. Looking at the boxplot of the students who did not receive psi, we observe that student with GPAs below 2.5 are not represented. Moreover, the boxplot is positioned lower compared to the others. To investigate the data further, we constructed histograms. We observe that for students who receive psi, the GPAs higher than 3.0 occur more frequently. In contrast, for students that did not receive psi, the GPAs between 2.5 and 3.0 occur more frequently. Hence, it can be argued that the data is biased because for the group of students who receive psi, the higher GPAs occur more frequently, whereas, for the group of students who do not receive psi, the lower GPAs occur more frequently.





Next, we investigate the relation between the variables passed and gpa. Looking at the boxplots, we clearly see that students who passed the test have higher GPAs and the students who did not pass the test have lower GPAs. The histogram confirms this by showing higher frequencies of higher GPAs for students that passed the test and higher frequencies of lower GPAs for students that did not pass the test. Hence, it could be argued that students who have a higher GPA are more likely to pass the test.

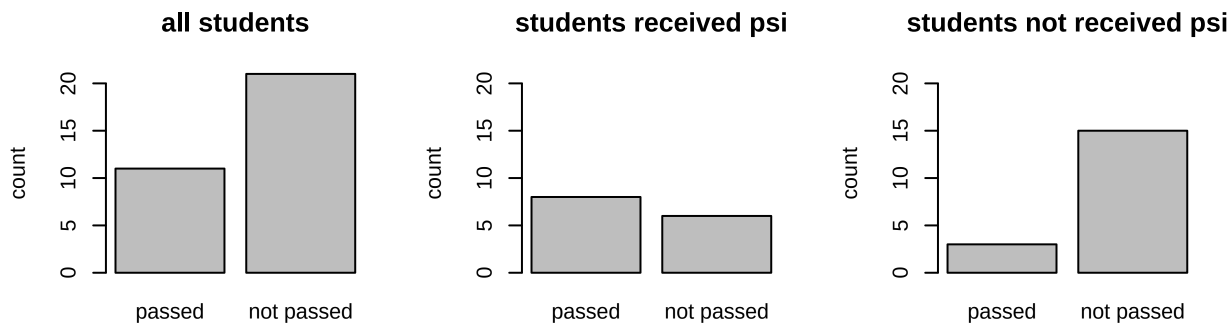


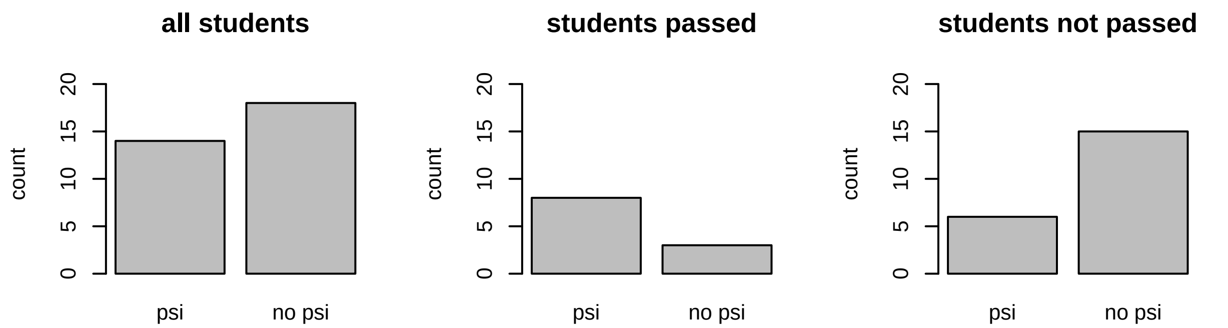


Lastly, we investigate the relation between the variables psi and passed. Looking at the barplots below, we observe that there are more students who did not pass the test compared to students who did pass the tests. In contrast, looking at the students who receiced psi, there are more students who passed than not passed, however, this difference is very small. For the students that did not receive psi, this difference is much larger and much more students did not pass compared to the students whi passed. When considering all the students again, we observe that the amount of students receiving and not receiving psi is evenly distributed. Slightly more students did not receive psi compared to the students who received psi. Moreover, we observe that of the students who passed, more received psi and of the students who did not pass, more did not receive psi.

The table shows the numbers for each possible combination of 'passed' and 'psi'. From this we could have a first look about the relation between this two variables.

```
##       psi
## passed  0  1
##      0 15  6
##      1  3  8
```





b)

We fit a logistic regression model that explains whether the psi will influence the passed. We test the null hypotheses that the psi do not influence passing the assignment. According to the summary below we could see that p-value for psi is smaller than significance level 0.05. Therefore, we reject H_0 here means psi works and will influence a student pass the assignment.

```
##
## Call:
## glm(formula = passed ~ psi + gpa, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8396  -0.6282  -0.3045   0.5629   2.0378
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -11.602      4.213  -2.754  0.00589 **
## psi            2.338      1.041   2.246  0.02470 *
## gpa            3.063      1.223   2.505  0.01224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 26.253  on 29  degrees of freedom
## AIC: 32.253
##
## Number of Fisher Scoring iterations: 5
```

c)

From the summary above, we could calculate the probability of a student who receives psi or not passing or not passing the assignment with gpa equal to 3. For student who receives the psi:

$$\frac{1}{1+e^{-(-11.602+2.338)+(3.063*3)}} = 0.481$$

For student who do not receives the psi:

$$\frac{1}{1+e^{-(-11.602+3.063*3)}} = 0.082$$

So the probability for student with gpa equal to 3 and receives psi to pass the assignment is 48.1%, while for student with gpa equal to 3 and not receives psi to pass the assignment is 8.2%.

d)

From the summary in b) we notice that the coefficient of psi is 2.338, which is positive, means rasing psi by 1 increases the linear predictor by 2.338 and increases the odds of passing the assignment by a factor e^2.338 which equal to 10.36049. This number means students who receive psi are 10.36049 times more likely to pass the assignment than those who do not receive psi. And this is not dependent on gpa as gpa and psi are independent to each other.

e) We test the null hypothesis p1 = p2. This means that the null hypotheses states that students who do not receive psi and students who receive psi show the same improvement. In the matrix we put the numbers 3, 15, 8 and 6. These number mean the following: from the 18 students who do not receive psi, 3 show improvement. This means that 18 - 3 = 15 students do not show improvements. From the 14 students who receive psi, 8 show improvement. This means that 14 - 8 = 6 students do not show improvement. Running Fisher's test when comparing the two binomial proportions, results in the p-value of 0.0265. This is smaller than the significance level of 0.05 and, therefore, H_0 is rejected. Therefore, we conclude that students receiving and not receiving psi do not show a similar improvement.

```
x=matrix(c(3,15,8,6),2,2); x
```

```
##      [,1] [,2]
## [1,]    3    8
## [2,]   15    6
```

```
fisher.test(x)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.0265
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02016297 0.95505763
## sample estimates:
## odds ratio
##  0.1605805
```

f) No, the second approach is not wrong, while it ignored the potential correlation between variables 'passed' and 'gpa'. But when considering whether a student receive psi or not is indenpendent to gpa, the second approach is correct.

g) Logistic regression: Advantages: it includes a predictive model which the Fisher exact test lacks. Disadvantages: it needs all explanatory variables to be independent to each other.

Fisher's test: Advantage: it is more suitable to work with smaller datasets as is the case in our experiment. Disadvantages: it is conservative and may be misleading.