STAT 1293 - Quiz 4

Gordon Lu

7/25/2020

Problem 1: ANOVA Test and Pairwise Comparison (20 points)

1a) Calculate the means and the standard deviations for the three groups. (4 points)

Solution:

```
tipping <- read.table("C:/Users/gordo/Desktop/tipping.txt", header = TRUE) #read in tipping.txt

tapply(tipping$Percent, tipping$Report, mean)
```

```
##    Bad   Good   None
## 18.180 22.220 18.725
```

```
tapply(tipping$Percent, tipping$Report, sd)
```

```
##      Bad     Good     None
## 2.098019 1.958947 2.388321
```

Yes, I think the standard deviations are roughly the same.

1b) Do the data support the hypothesis that there are differences among the tipping percentages for the three experimental conditions? Conduct an ANOVA F test. (8 points)

Solution:

```
anova(lm(tipping$Percent ~ tipping$Report))
```

```
## Analysis of Variance Table
##
## Response: tipping$Percent
##                Df Sum Sq Mean Sq F value    Pr(>F)
## tipping$Report  2 192.22  96.112  20.679 1.767e-07 ***
## Residuals      57 264.92   4.648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4-Step H.T.

Hypothesis: $H_0 : \mu_{good} = \mu_{bad} = \mu_{none}$

H_a: Not all the μ_i's are equal. $\mu_i \neq \mu_j$ for some i, $j \in good, bad, none$ and $i \neq j$

F statistic: $F_0 = \frac{MSG}{MSE} =$ `20.679`

P-value: $P(F_{I-1,N-I} > F_0) =$ `1.767e-07`

Decision: Since the p-value is 1.767e-07, we reject the null at the significance level `0.01`, and conclude that not all the means are equal.

Conclusion: In other words, we have sufficient evidence to conclude that there is a difference among the tipping percentages for the three experimental conditions.

1c) Where does the true difference lie? Use a pairwise comparison to answer this question. (4 points)

Solution:

```
#By using "Bad" as the ref level, can easily compare to Good and None
tipping$Report <- relevel(tipping$Report, ref = "Bad")
summary(lm(tipping$Percent~tipping$Report, data = tipping))
```

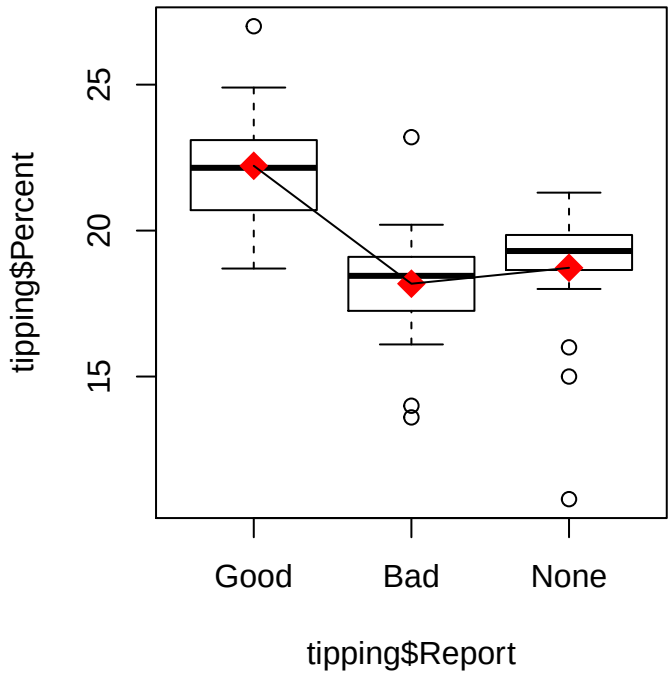
```
##
## Call:
## lm(formula = tipping$Percent ~ tipping$Report, data = tipping)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9250 -0.7988  0.3475  1.0337  5.0200
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          18.1800     0.4821  37.713  < 2e-16 ***
## tipping$ReportGood    4.0400     0.6817   5.926  1.9e-07 ***
## tipping$ReportNone    0.5450     0.6817   0.799    0.427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.156 on 57 degrees of freedom
## Multiple R-squared:  0.4205, Adjusted R-squared:  0.4002
## F-statistic: 20.68 on 2 and 57 DF,  p-value: 1.767e-07
```

The true differences lie between "Bad" and "Good". Note that "Bad" and "None" are not significantly different, while "Bad" and "Good" are significantly different.

1d) Use a side-by-side boxplot with means to show the relationship between the means. You don't have to reorder the levels. (4 points)

Solution:

```
par(pty = "s")
tipping$Report = factor(tipping$Report, levels = levels(tipping$Report)[c(2,1,3)])
boxplot(tipping$Percent ~ tipping$Report, data = tipping)
means <- tapply(tipping$Percent, tipping$Report, mean)
points(means, col="red", pch = 18, cex = 2)
lines(1:3, means, cex = 2)
```



Problem 2: Inference on one proportion (15 points)

2a) Are the conditions of large-sample procedure satisfied? Show your evidence. (3 points)

Solution:

```
#First let's confirm that we have binary outcomes:

#Them confirm that the trials are independent:

#Then confirm that the sampling distribution is normally distribution.

#We also know the data is from a SRS
p_success <- 19/172
p_failure <- 1 - p_success
```

```r
172 * p_success
```

```
## [1] 19
```

```r
172 * p_failure
```

```
## [1] 153
```

Yes. First observing that our data comes from a SRS, we can assume the outcomes are unbiased and random. Additionally, notice how each individual's response can be summarized with a binary outcome. That being, either "Yes" or "No". Then, note how the response of one individual does not impact the response of another, thus individuals are independent. Lastly, note that $np \geq 10 = 19$ and $n(1-p) \geq 10 = 153$. Thus, we can conclude that the sample size if sufficiently large, and importantly that the sampling distribution of p is normally distributed.

2b) Calculate a 99% confidence interval for the proportion of all undergraduates at this university who would report cheating. Use correct=F. (4 points)

Solution:

```r
prop.test(19, 172, correct = F, conf.level = 0.99)$conf.int
```

```
## [1] 0.06281231 0.18705431
## attr(,"conf.level")
## [1] 0.99
```

Carefully interpret the confidence interval you got in Part (a). (2 point)

We can say with 99% confidence that the proportion of undergraduates at the university who would report cheating is at the very least approximately 6.28%, and at the very most approximately 18.71%.

2c) Does the data provide sufficient evidence to claim that less than 20% undergraduates in the university would report cheating? Use correct=F. (8 points)

Solution:

```r
prop.test(19, 172, 0.20, correct = F, alt = "less")
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  19 out of 172, null probability 0.2
## X-squared = 8.6177, df = 1, p-value = 0.001665
## alternative hypothesis: true p is less than 0.2
## 95 percent confidence interval:
##  0.0000000 0.1559705
## sample estimates:
##         p
## 0.1104651
```

4-Step H.T.

Hypothesis: $H_0 : p = 0.20$

$H_a : p < 0.20$

z statistic: $z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 8.6177$

P-value: $P(Z < z_0) = 0.001665$

Decision: Since the p-value is 0.001665, we reject the null at the significance level 0.01, and conclude that there is sufficient evidence to conclude that less than 20% undergraduates in the university would report cheating.

Conclusion: In other words, we have sufficient evidence to conclude that less than 20% undergraduates in the university would report cheating.

Problem 3: Inference on two proportions (15 points)

3a) Are the conditions of large-sample inference satisfied ? Show your evidence. (4 points)

Solution:

```
p1_success <- 2818/3239
p1_failure <- 421/3239

3239*p1_success
```

```
## [1] 2818
```

```
3239*p1_failure
```

```
## [1] 421
```

```
p2_success <- 2091/2787
p2_failure <- 696/2787

2787*p2_success
```

```
## [1] 2091
```

```
2787*p2_failure
```

```
## [1] 696
```

Yes. We have the following populations: Those who received formal music instruction during all three years, which is 3239, and those who did not receive formal music instruction during all three years of middle school is 2787. Now, for the 3239 ninth graders who received formal musical instruction, 2818 passed algebra, meaning the probability of success for the first population is $\frac{2818}{3239}$. It is important to note that whether or not a student has passed algebra does not impact another student's result. Therefore for the first population, the students are independent of one another. Additionally, notice that $\frac{2818}{3239} \times 3239 = 2818 \geq 10$, and

$(1 - \frac{2818}{3239}) \times 3239 = 421 \geq 10$, therefore the sample sizes for population 1 is sufficiently large, and thus the sampling distribution for population 2 is approximately normally distributed. The same approach can be used for population 2. Note that among the 2787 students who did not receive formal music instruction, 2091 passed algebra, meaning the probability for the second population is $\frac{2091}{2787}$. It is also important to note that whether or not a student has passed algebra does not impact another student's result. Therefore, for the second population, the students are independent of one another as well. Additionally, notice that $\frac{2091}{2787} \times 2787 = 2091 \geq 10$, and $(1 - \frac{2091}{2787}) \times 2787 = 696 \geq 10$, therefore the sample size for population 2 is suffiently large, and thus we can conclude the sampling distribution for population 2 is approximately normally distributed.

3b) Calculate a common sample proportion ($\hat{p}$) of receiving a passing grade for all students in the sample. (1 point)

Solution:

```
p_hat <- (2818 + 2091)/6026
p_hat
```

```
## [1] 0.8146366
```

3c) Calculate the difference between the sample proportions of the two groups. (1 point)

Solution:

```
p_hat1 <- 2818/3239
p_hat2 <- 2091/2787

diff <- p_hat1 - p_hat2
diff
```

```
## [1] 0.1197525
```

3d) Calculate a 95% confidence interval for the difference between the proportions of students receiving a passing grade. Use correct=F. (3 points)

Solution:

```
prop.test(c(2818,2091), c(3239,2787),correct=F)$conf.int
```

```
## [1] 0.09994411 0.13956090
## attr(,"conf.level")
## [1] 0.95
```

3e) Does the data provide sufficient evidence to claim that musical instruction helps in passing an algebra class? Conduct a H.T. at 0.05 significance level. Use correct=F. (6 points)

Solution:

```
prop.test(c(2818,2091), c(3239,2787), correct=F, alt = "greater", conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(2818, 2091) out of c(3239, 2787)
## X-squared = 142.27, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1031288 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.8700216 0.7502691
```

4-Step H.T.

Hypothesis: $H_0 : p_{pass-and-play-instrument} = p_{pass-and-not-play-instrument}$

$H_a : p_{pass-and-play-instrument} > p_{pass-and-not-play-instrument}$

z statistic: $z_0 = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}} =$ `142.27`

P-value: $P(Z > z_0) <$ `2.2e-16`

Decision: Since p < 0.05, we have sufficient evidence to reject the null hypothesis.

Conclusion: Therefore, we have sufficient evidence to conclude that musical instruction helps in passing algebra.