Descriptive Statistics of a Survey of Seniors of the Class of 2018 and Faculty in the JBHS Class of 2018

AP Statistics Second Quarter Project

Benjamin El-Wardany

Rahul Kiefer

Jonathan Lam

PART 1

COMPLETED IN CLASS

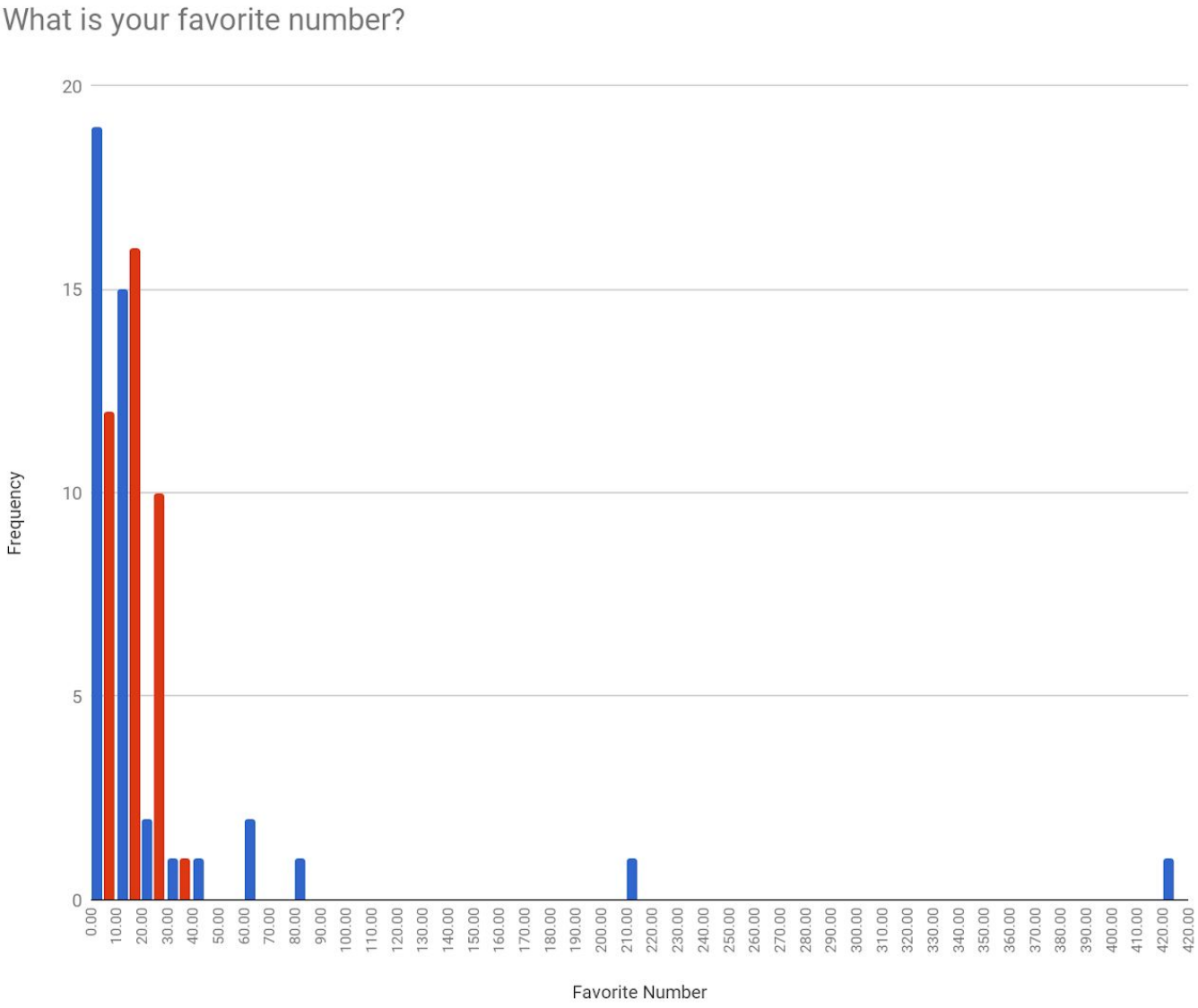
PART 2

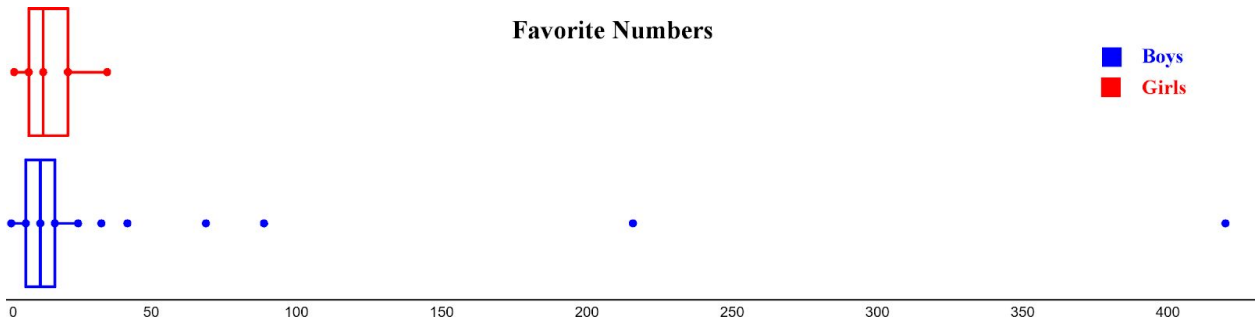
TASK 1: QUANTITATIVE VARIABLE

What is your favorite number? (frequency table)		
Number	Frequency	
	Males	Females
2	2	0
3	1	1
4	4	2
6	1	2
7	5	3
8	4	3
9	2	1
10*	1	0
11	1	3
12	3	2
13	2	3
14	1	2
15	1	0
16	2	5
17	3	0
18	1	1
21	1	1

22	0	3
23	0	1
25	1	0
26	0	1
27	0	3
29	0	1
33	1	0
35	0	1
42	1	0
69	2	0
89	1	0
216	1	0
420	1	0

* interpreted from: “Well, on Saturday it is 10 because I like that number, but on most days of the week is 1 cause Im the best ever. No one can stop me. But overall my favorite number is 10, so yes 10.”


What is your favorite number?
20
15
10
5
0
Frequency
0.00
10.00
20.00
30.00
40.00
50.00
60.00
70.00
80.00
90.00
100.00
110.00
120.00
130.00
140.00
150.00
160.00
170.00
180.00
190.00
200.00
210.00
220.00
230.00
240.00
250.00
260.00
270.00
280.00
290.00
300.00
310.00
320.00
330.00
340.00
350.00
360.00
370.00
380.00
390.00
400.00
410.00
420.00
420.00
Favorite Number


Favorite Numbers
Boys
Girls
0
50
100
150
200
250
300
350
400

Numerical Summary

	Min.	Q1	Med.	Q3	Max.	IQR	$\bar{x}$	S_x
Males	2	7	12	17	420	10	30.4884	70.4311
Females	3	8	13	21.5	35	13.5	14.8718	7.9676

Outlier test:

Males: outliers at 33, 42, 69, 69, 89, 216, 420

Lower fence: Q_1 - 1.5IQR = 7 - 1.5(10) = -8

Upper fence: Q_3 + 1.5IQR = 17 + 1.5(10) = 32

Females: no outliers

Lower fence: Q_1 - 1.5IQR = 8 - 1.5(13.5) = -12.25

Upper fence: Q_3 + 1.5(IQR) = 21.5 + 1.5(13.5) = 41.75

Because the distribution of the males and females differed greatly in some of their features, we decided to analyze each one separately.

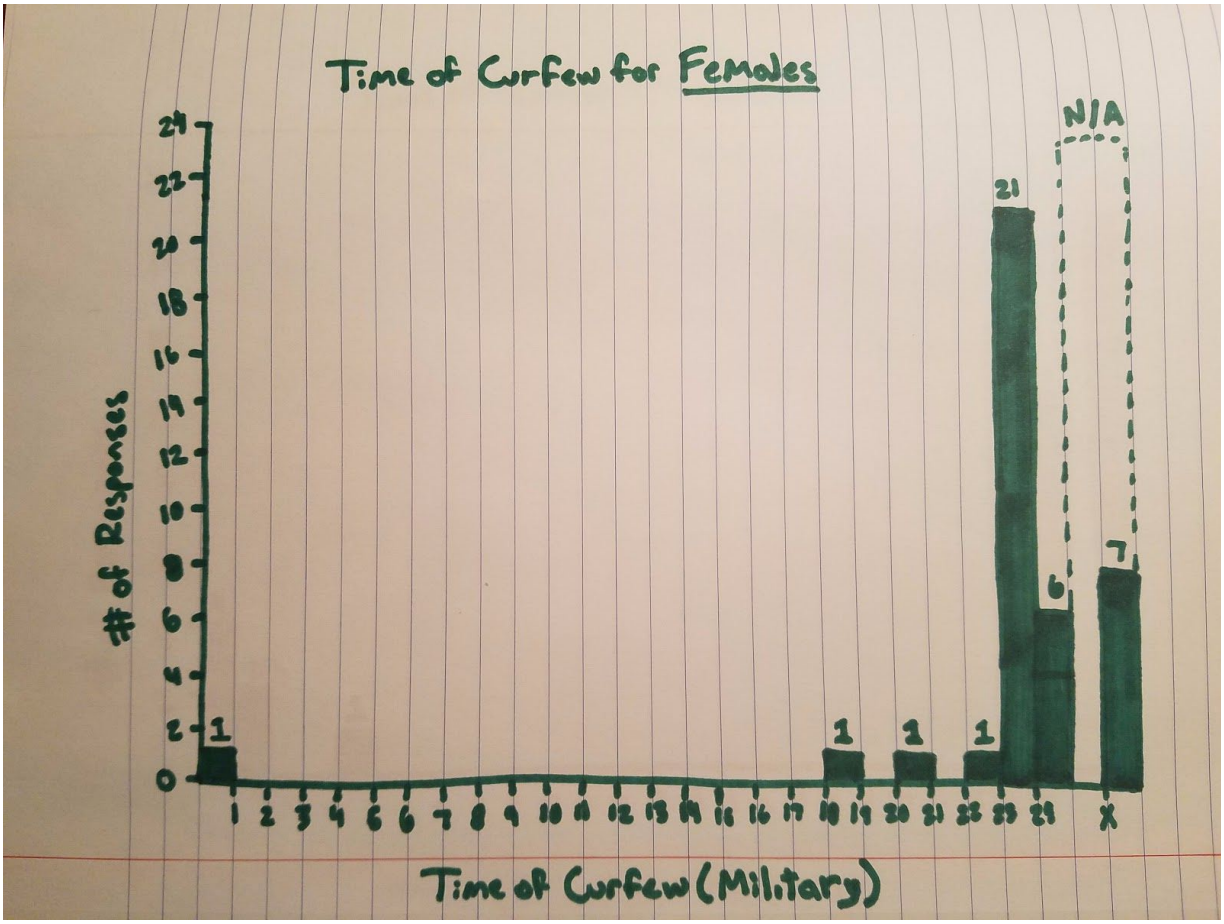
The shape of the distribution of favorite numbers is heavily skewed right and unimodal, with a mode at 0 to 9 for males and a mode at 10 to 19 for females. The median for males is 12, which is similar to the females' median of 13. Both demonstrated a small IQR (10 for males and 13.5 for females), but there was a great difference in range. For females, there were no outliers nor gaps and the range was 33. For males, there were 7 high outliers, including 6 high outliers, drastically increasing the mean and standard deviation and making the graph much more asymmetric. The range is 418 (13 times as great as that of the females). Thus, while the middle 50% of both distributions are tightly clustered around the 0-19 numbers, the distribution for males has some very high outliers and a great spread, while the spread and range of females is relatively small.

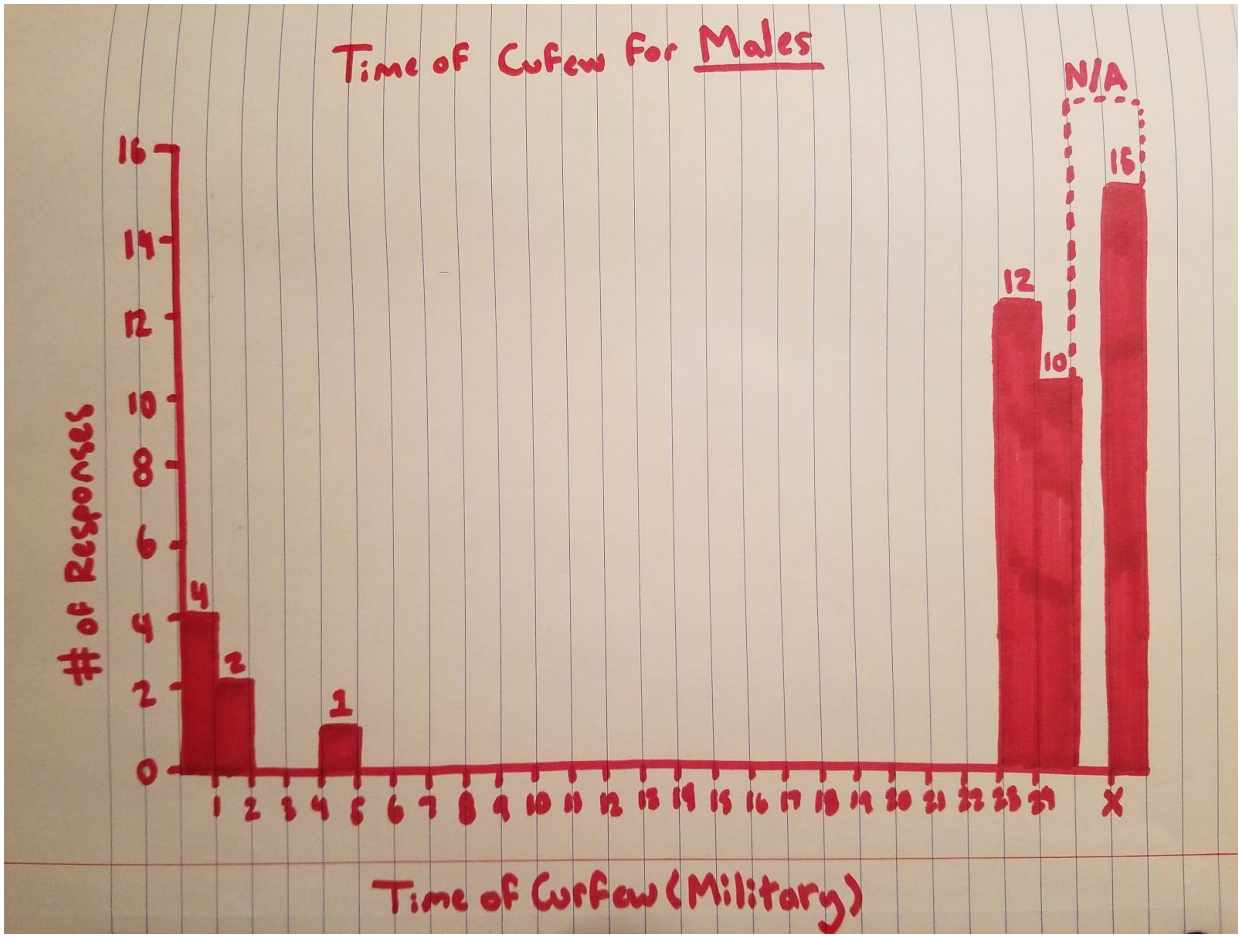
TASK 2: CATEGORICAL VARIABLE

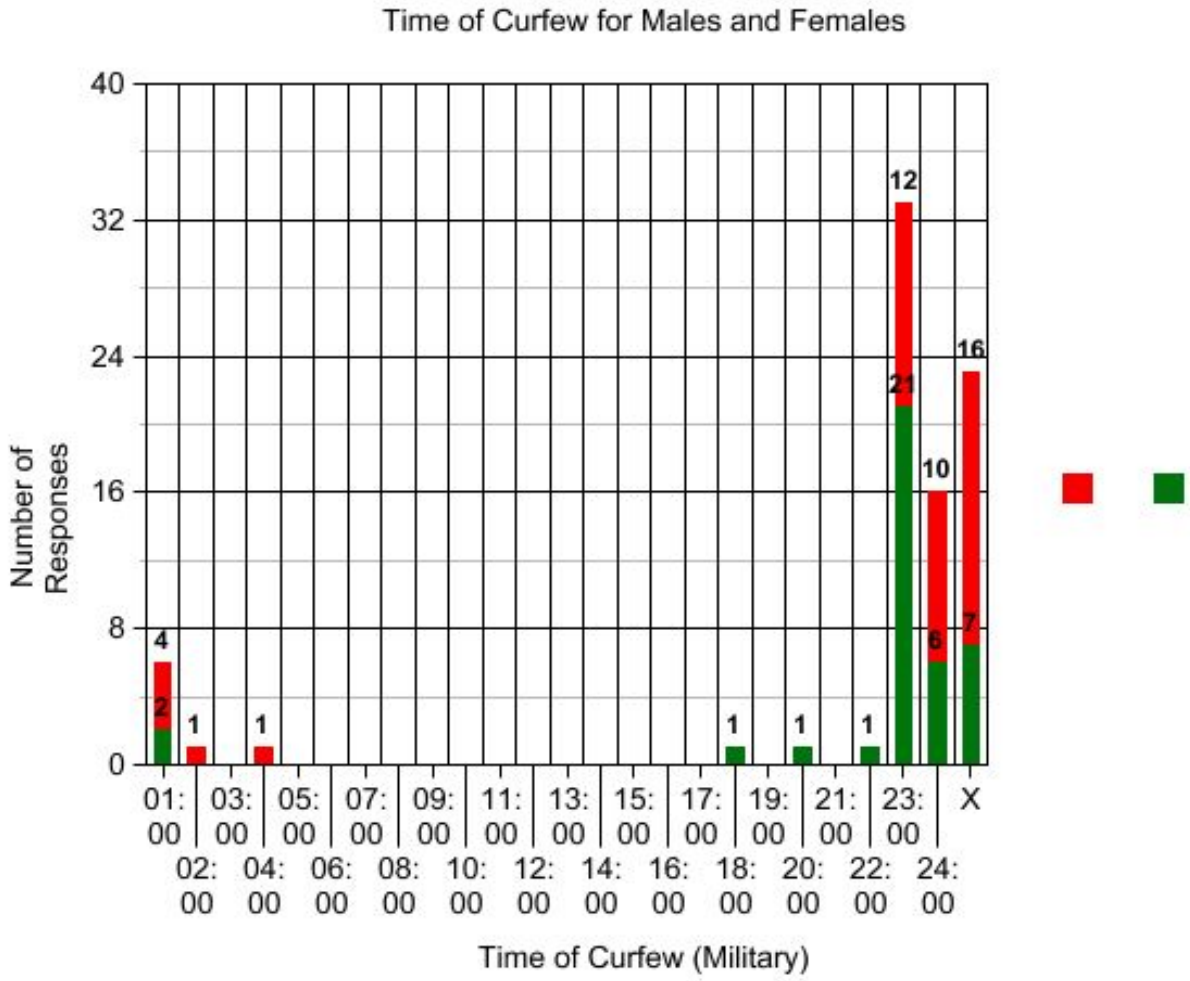
Curfew on weekends (responses)			
Males		**Females**	
response	**interpreted**	**response**	**interpreted**
12 p.m.	24:00	"whenever I want"	X
"12 at nightime"	24:00	"no curfew"	X
"none"	X	"11 if I'm driving"	23:00
11	23:00	11:00 PM	23:00
11	23:00	10:00	22:00
"no curfew"	X	11:00	23:00
1:30 a.m.	1:00	11:00	23:00

"N/A"	X	11:00	23:00
12:00-12:30	1:00	1	1:00
11:00	23:00	12-1	1:00
12:00	24:00	11	23:00
12	24:00	8	20:00
11:30	23:00	11:30	23:30
"Don't have one"	X	11	23:00
1	1:00	11	23:00
"none"	X	11	23:00
"none"	X	12:30	1:00
"don't know"	X	11	23:00
"none"	X	11:00 PM	23:00
"none"	X	n/a	X
2:00 AM	2:00	11:00 PM	23:00
11:00 PM	23:00	12:00 AM	24:00
11	23:00	12:00 AM	24:00
11	23:00	6 (am/pm?)	18:00
12	24:00	12:00 AM	24:00
"when I want"	X	11:00 PM	23:00
2	2:00	11pm (CT law) n/a otherwise	23:00
"dont have one"	X	11	23:00
12	24:00	11	23:00
NA	X	"don't have one"	X
1	1:00	11	23:00
11:30	23:00	11	23:00
≈12	24:00	none	X
12ish	24:00	don't have one	X
none	X	11:30	23:00
1	1:00	11	23:00
11	23:00	"don't have one"	X
4	4:00	12	24:00

11	23:00		
12	24:00		
11	23:00		
"never"	X		
N/A	X		
11	23:00		




Time of Curfew for Males
N/A
# of Responses
Time of Curfew (Military)
X


Time of Curfew for Males and Females
Number of Responses
40
32
24
16
8
0
4
2
1
1
1
1
1
12
21
10
6
16
7
01: 00
02: 00
03: 00
04: 00
05: 00
06: 00
07: 00
08: 00
09: 00
10: 00
11: 00
12: 00
13: 00
14: 00
15: 00
16: 00
17: 00
18: 00
19: 00
20: 00
21: 00
22: 00
23: 00
24: 00
X
Time of Curfew (Military)

Chi-Squared Test For Independence

Hypotheses:

- Null: Gender and Curfew Time are independent of each other
- Alternate: Gender and Curfew Time are not independent of each other

Conditions:

- Counted Data Condition: There are counts for individuals for 2 categorical variables: Curfew Time and Gender
- Randomization: We were given randomly selected students to survey so we can believe that the sample is representative of the Population
- Expected Cell Frequency:

4	1	2.71	2.2
2	0	1.08	.91
1	0	.54	.45
0	1	.54	.45
0	1	.54	.45
0	1	.54	.45
12	21	17.89	15.11
10	6	8.67	7.33
16	7	12.47	10.53

$\Rightarrow$

There are not at least 5 expected values in each cell

Although the Expected Cell Frequency condition is not met, we will proceed to the Chi-Squared test with a df of 8

Mechanics:

$$\frac{(4-2.71)^2}{2.71} + \frac{(1-2.2)^2}{2.2} + \dots \frac{(7-10.53)^2}{10.53}$$

$\chi^2 = 14.2875$

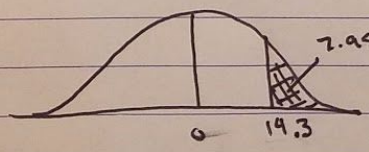
$p = .07957$

$df = 8$

$df = (r-1)(c-1)$

$(9-1)(2-1)$

$(8)(1) = 8$



P-value $= P(\chi^2 > 14.2875) \approx .07957$

Conclusions

Since our P-value of .07957 is less than .1, we reject the null hypothesis that Gender and Curfew Time are independent of each other. Therefore we have evidence to support the Alternate Hypothesis, which states that Gender and Curfew Time are not independent of each other.

TASK 3: AGE OF CARS COMPARISON

- Change statistics from "year of car" to "age of car" (2018 - year of car)
- Make back-to-back stem and leaf plot

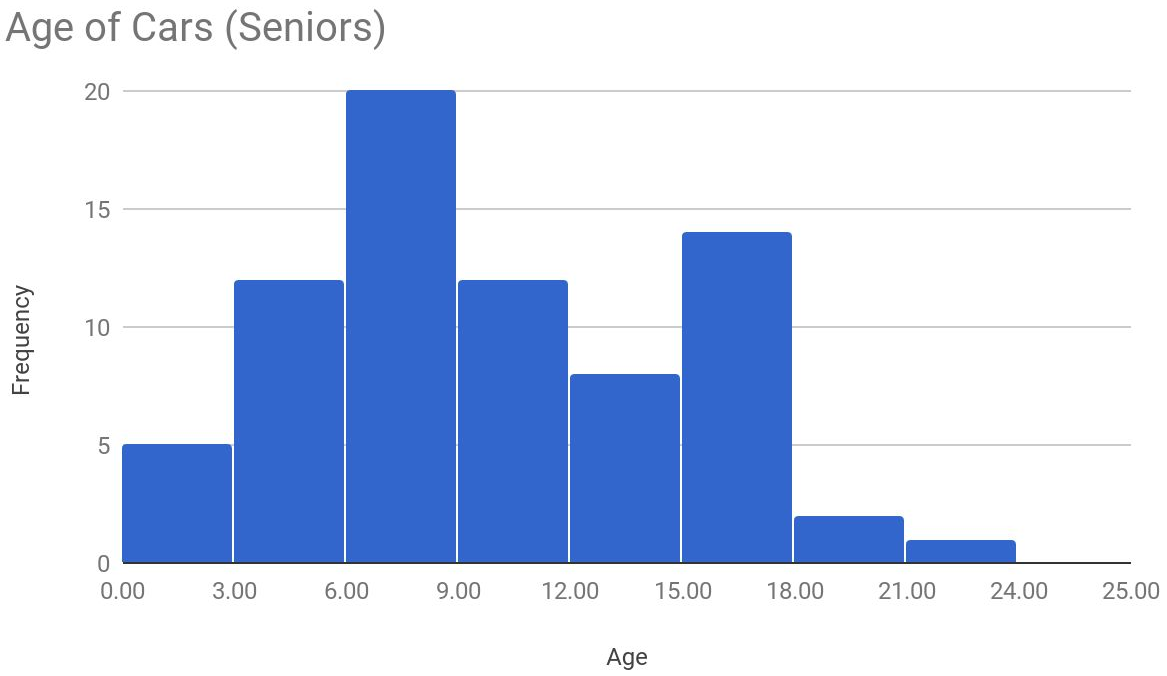
Ages of Cars Back-to-Back Stem and Leaf Plot

Students		*Staff*
4444443332211 1	0	0111111112222222222233333333333444444444444
99888888777776666666666555	0	555555556666666677778899
444432221111100000	1	00011111111111333334
9877777666666655	1	5569
3	2	2

Key: **1 | 0** represents a 2008 model car*

74 student responses were usable, 12 were unusable ("N/A" or "Don't drive"), 10 were nonresponse. 91 staff responses were recorded (all usable), 54 were nonresponse.

* age of car is calculated by 2018 minus the car's make year. Responses that were a range of years (e.g., 2015-2016) or half-years (e.g. 2016.5) were rounded down before subtracting from 2018.



Numerical Summary

	Min.	Q1	Med.	Q3	Max.	IQR	$\bar{x}$	S_x
Students	1	6	8.5	14	23	8	9.5541	5.1820
Staff	0	3	5	10.5	22	7.5	6.2956	4.6113

Outlier test:

Students: No outliers

Lower fence: Q_1 - 1.5IQR = 6 - 1.5(8) = -6

Upper fence: Q_3 + 1.5IQR = 14 + 1.5(8) = 26

Staff: Outlier at 22

Lower fence: Q_1 - 1.5IQR = 3 - 1.5(7.5) = -8.25

Upper fence: Q_3 + 1.5(IQR) = 10.5 + 1.5(7.5) = 21.75

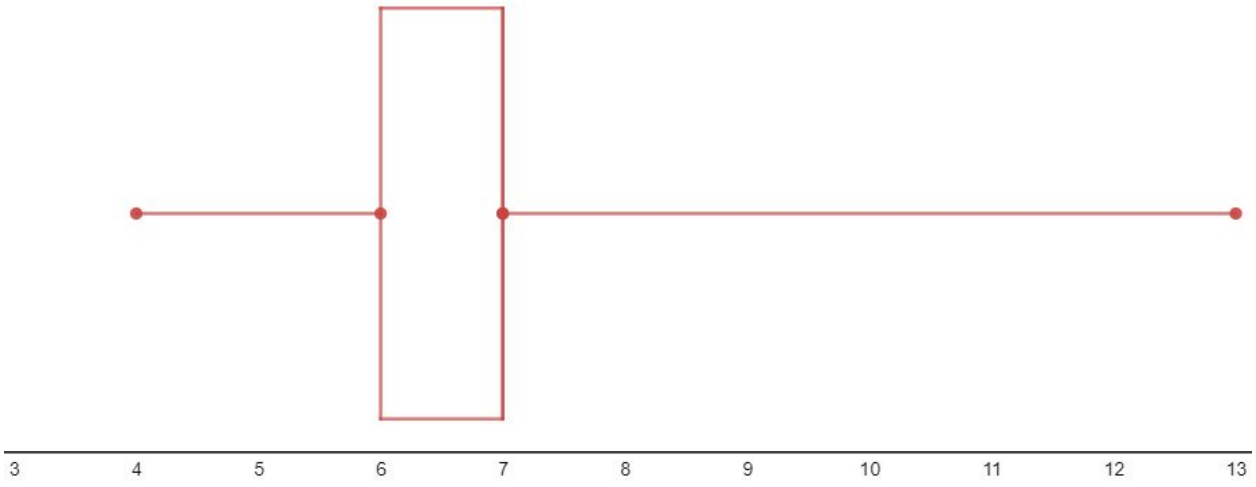
Both the distribution of ages of cars driven by students and staff to school are skewed right and bimodal. The ages of cars held by students has a large peak at 6 to 8 years old, and a smaller peak at 15 to 17 years old. The modes for the ages of cars driven by staff are farther left, with a larger mode at 3 to 5 years old and 9 to 11 years old, and it is skewed more strongly right. The median of the ages of the cars students drive to school is greater than that of the staff, 8.5 years old on average compared to 5 years old. The spread of the car age distributions are very similar: the IQR of the students' cars' ages is 8 years, as compared to 7.5 years for the staff, and the minimum and maximum ages for the students are only one year apart from those of the staff. There are no gaps in either distribution, no outliers in the students'

distribution, and only one outlier out of the staff's distribution at 22 years old. Overall, the distributions of the ages of the cars students and staff drive to school have a similar shape (skewed right with two differently-sized modes) and similar range (middle 50% lies within 7.5 to 8 years), but the distribution of the staff is shifted left from that of the students. In other words, the average age of a staff's car is smaller than that of a student's by approximately 3.5 years.

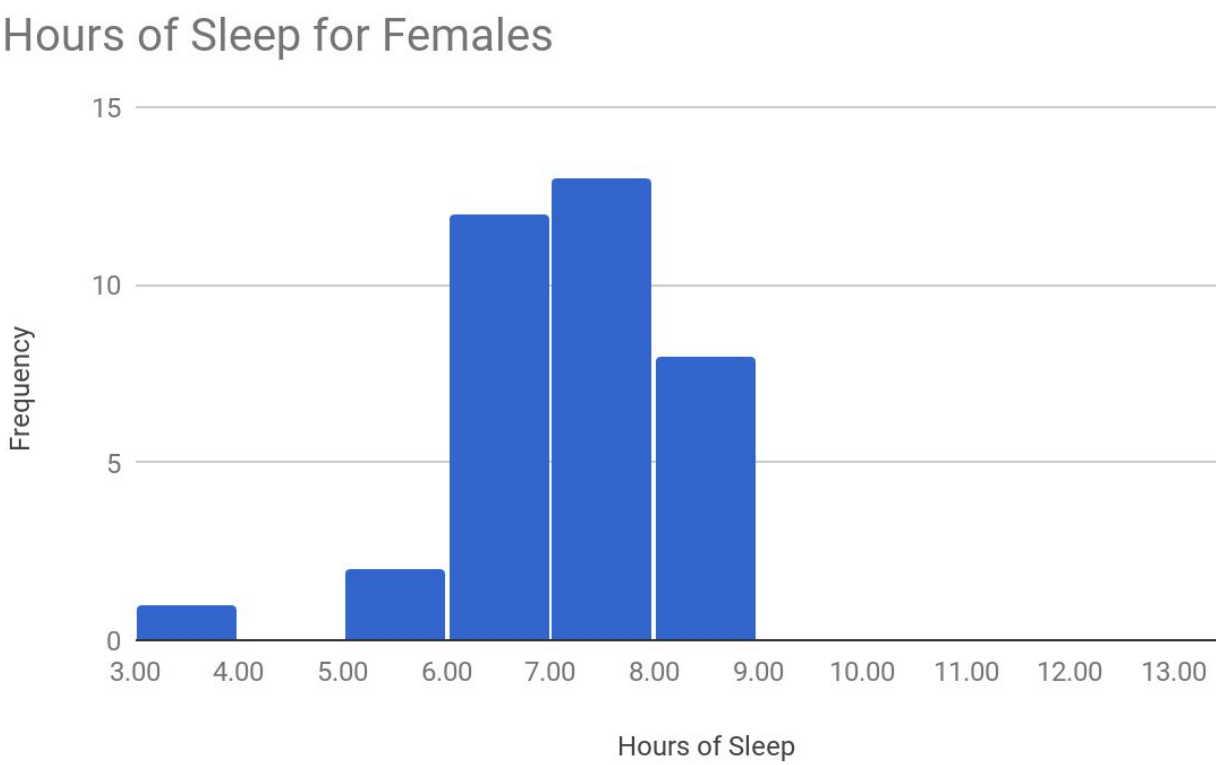
TASK 4: QUANTITATIVE AND CATEGORICAL VARIABLE

Hours of Sleep Per School Night (frequency table)		
	Frequency	
Response (hours)	**Males**	**Females**
3.5	0	1
4	2	0
5	2	2
6	10	12
6.5	5	0
7	15	12
7.5	0	1
8	8	7
8.5	0	1
8.8	1	0
13	1	0

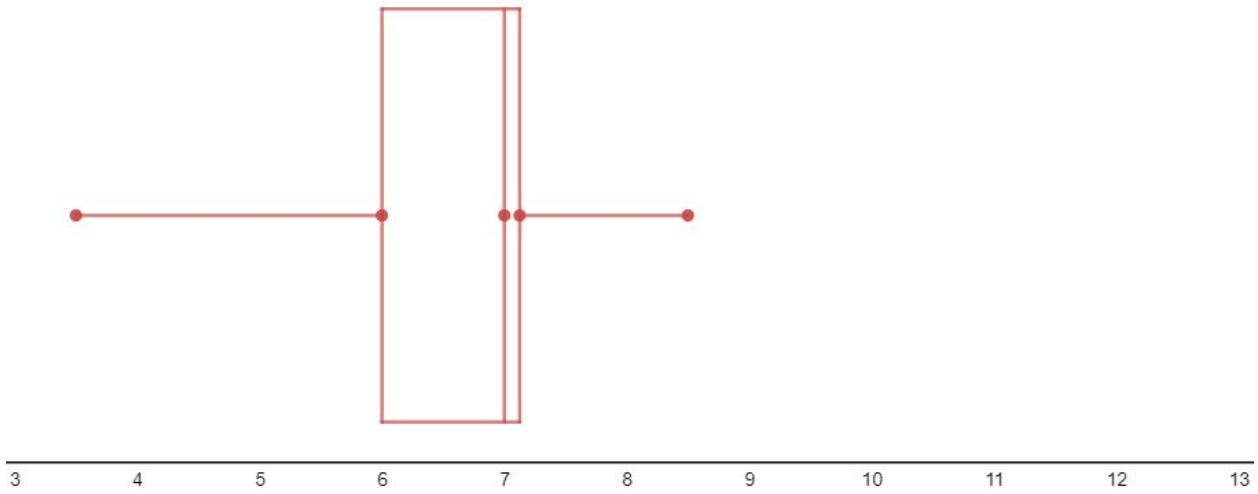
Box Plot for Males



Hours of Sleep



Box Plot for Females



Numerical Summary

	Min.	Q1	Med.	Q3	Max.	IQR	$\overline{x}$	S_x
Males	4	6	7	7	13	1	6.8477	1.3933

Females	3.5	6	7	7	8.5	1	6.6711	1.0350

Outlier test:

Males: outliers at 4, 4, 8.8, 13 hours

Lower fence: Q_1 - 1.5IQR = 6 - 1.5(1) = 4.5

Upper fence: Q_3 + 1.5IQR = 7 + 1.4(1) = 8.5

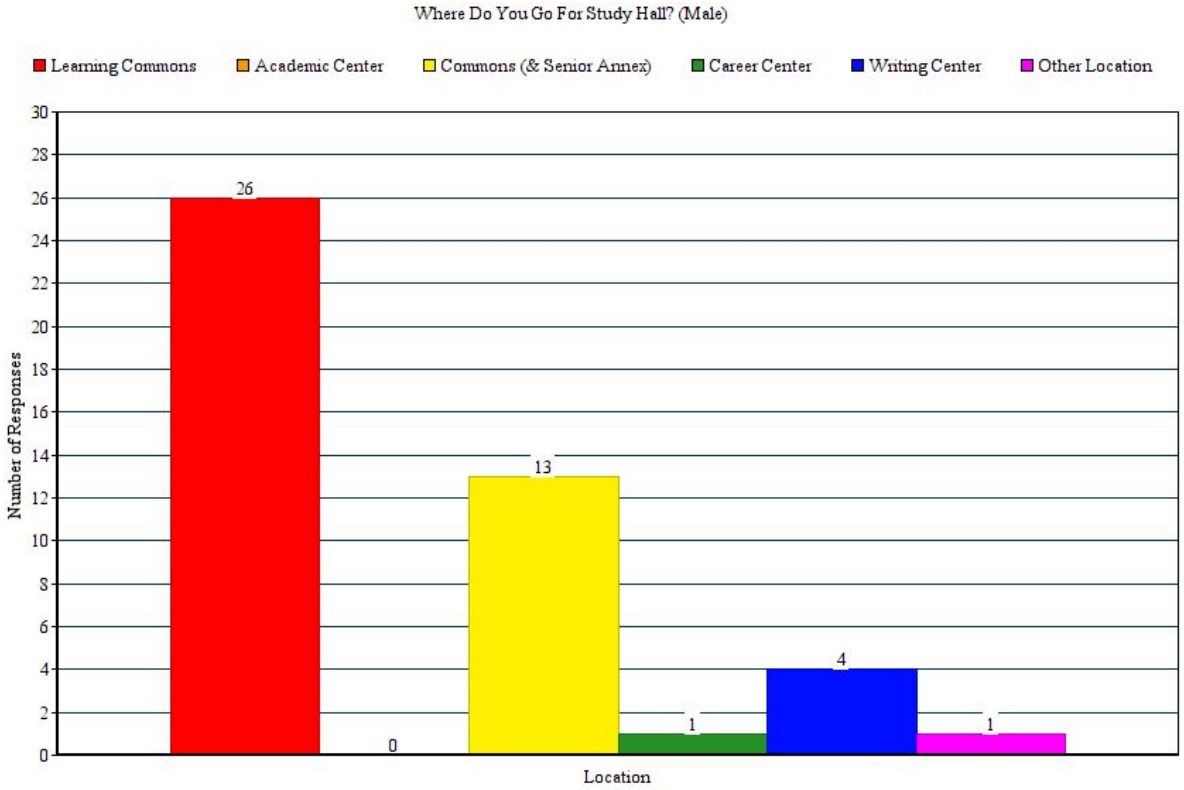
Females: Outlier at 3 hours

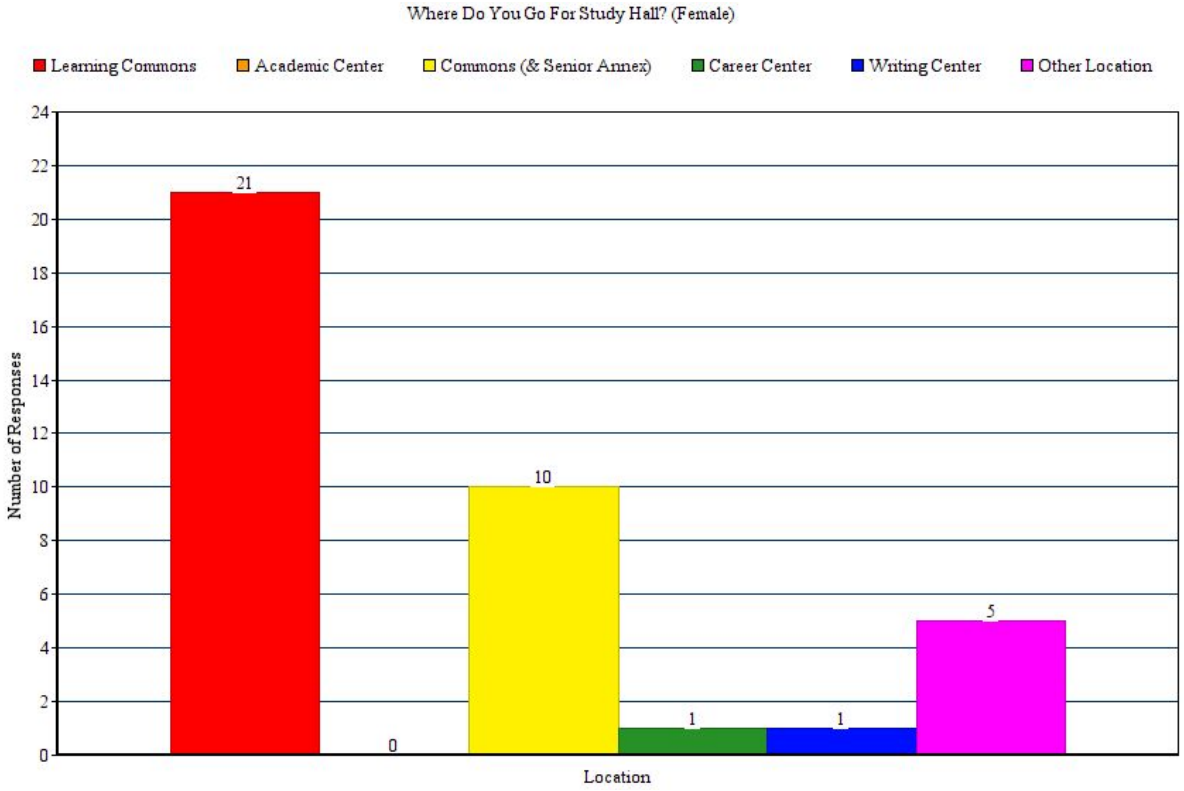
Lower fence: Q_1 - 1.5IQR = 6 - 1.5(1) = 4.5

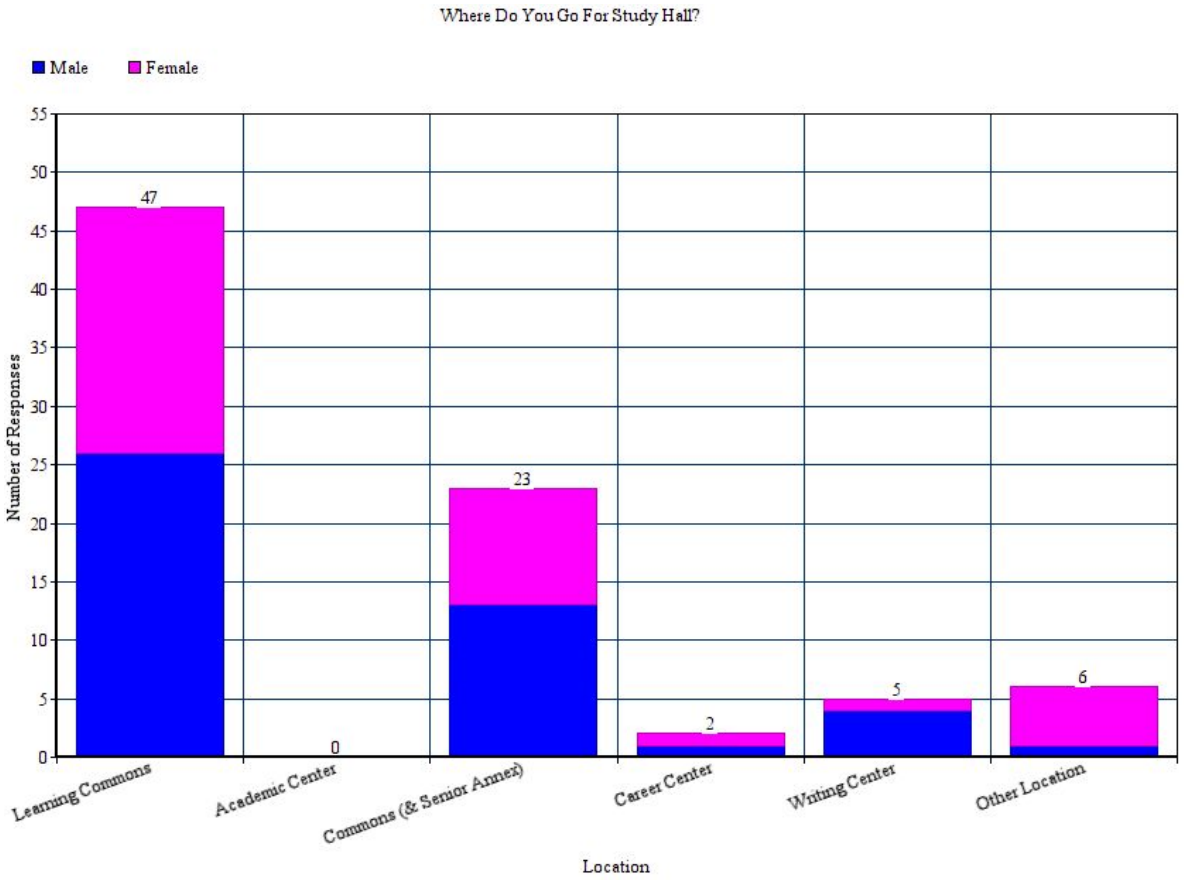
Upper fence: Q_3 + 1.5(IQR) = 7 + 1.5(1) = 8.5

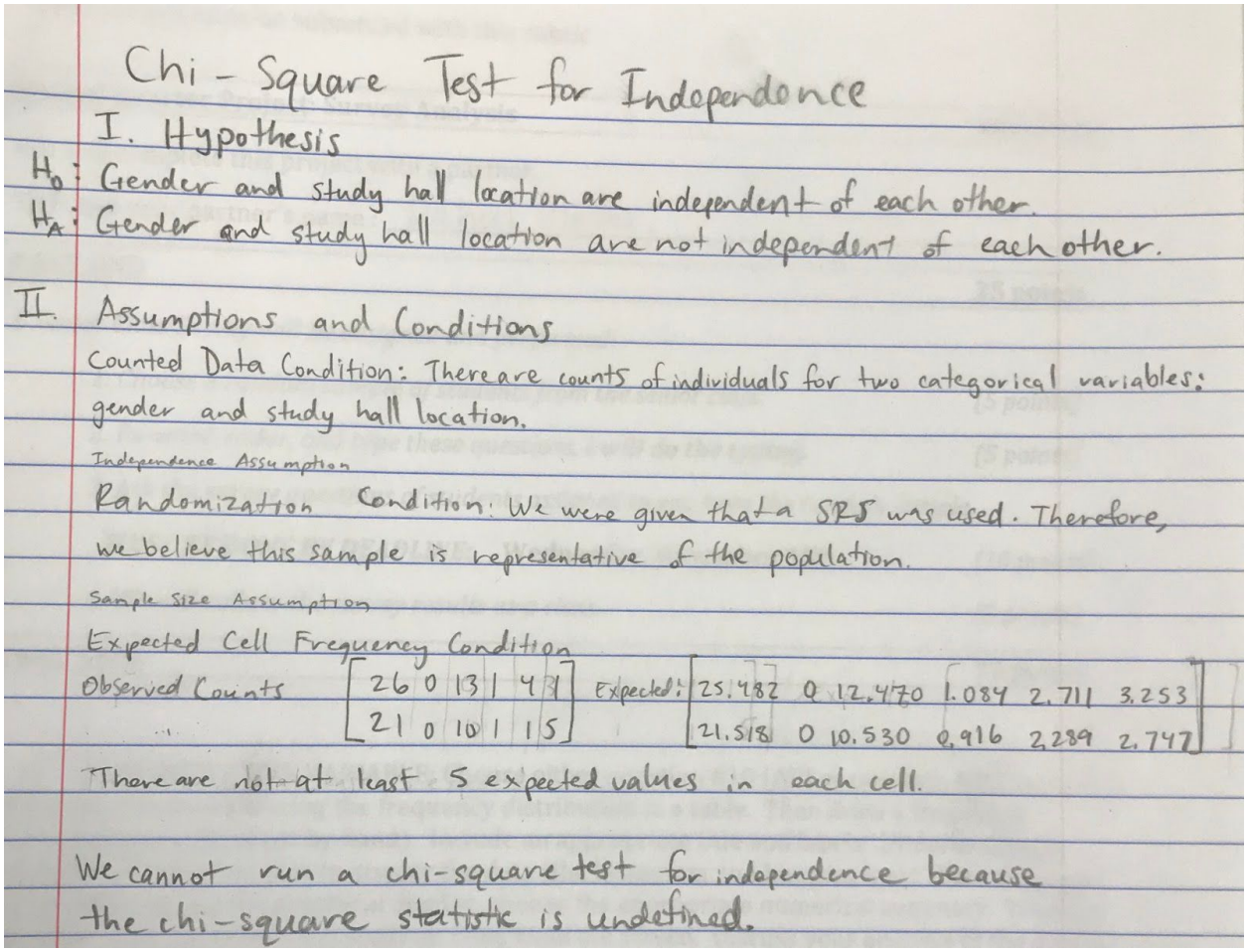
For males, the distribution of hours of sleep is unimodal with a mode at 7 and is skewed to the right, while for females the distribution is unimodal with a mode at 7 and skewed to the left. For males, the median is 7 hours with an IQR of 1, and for females, the median and IQR are the same. There are outliers of 4, 8.8, and 13 hours for males; and of 3 hours for females (see outlier tests above). When we disregard the outliers, the median values for hours of sleep for both males and females are equal (as well as Q3 values), suggesting that the number of hours of sleep for someone should not depend on their gender.

Where Do You Go For Study Hall? (two-way table)						
	Learning Commons	**Academic Center**	**Commons (& Senior Annex)**	**Career Center**	**Writing Center**	**Other Location**
Males	26	0	13	1	4	1
Females	21	0	10	1	1	5


Where Do You Go For Study Hall? (Male)
Learning Commons
Academic Center
Commons (& Senior Annex)
Career Center
Writing Center
Other Location
Number of Responses
Location
26
0
13
1
4
1


Where Do You Go For Study Hall? (Female)
Learning Commons
Academic Center
Commons (& Senior Annex)
Career Center
Writing Center
Other Location
Number of Responses
Location
21
0
10
1
1
5


Where Do You Go For Study Hall?
Male
Female
Number of Responses
55
50
45
40
35
30
25
20
15
10
5
0
47
0
23
2
5
6
Learning Commons
Academic Center
Commons (& Senior Annex)
Career Center
Writing Center
Other Location
Location


Chi - Square Test for Independence

I. Hypothesis

H_0: Gender and study hall location are independent of each other.

H_A: Gender and study hall location are not independent of each other.

II. Assumptions and Conditions

Counted Data Condition: There are counts of individuals for two categorical variables: gender and study hall location.

Independence Assumption

Randomization Condition: We were given that a SRS was used. Therefore, we believe this sample is representative of the population.

Sample Size Assumption

Expected Cell Frequency Condition

Observed Counts $\begin{bmatrix} 26 & 0 & 13 & 1 & 4 & 1 \\ 21 & 0 & 10 & 1 & 1 & 5 \end{bmatrix}$ Expected: $\begin{bmatrix} 25.482 & 0 & 12.470 & 1.084 & 2.711 & 3.253 \\ 21.518 & 0 & 10.530 & 0.916 & 2.289 & 2.747 \end{bmatrix}$

There are not at least 5 expected values in each cell.

We cannot run a chi-square test for independence because the chi-square statistic is undefined.

TASK 5: SURVEY TECHNIQUE ANALYSIS

Reducing selection and response biases

In order to reduce selection bias, we selected students from a list of all of the JBHS 2018 seniors using an SRS. Because all of the students from the senior class were from the list that the SRS, no student had a different chance of being selected than any other.

To reduce response bias, we reviewed the questions in the survey before it was sent out to try and change any language that we thought favored one answer choice over another. Students were also instructed to inform the survey respondents that they were allowed to leave their surveys anonymous and submit the answers directly to Mrs. Sopko's mailbox so that the surveyor could not see his or her responses. These measures were taken to avoid giving the responder any embarrassment or other pressure that would cause them to provide false responses.

Despite our efforts to reduce response bias, there was still problems with the wording of the survey that was sent out that caused inconsistencies between responses. For example, one of the responses to the question, "How many schools did you apply early?" was: "all." We should have specified that a numerical answer should have been inputted. Similarly, one answer for the question "What is your favorite number?" was: "Well, on Saturday it is 10 because I like that number, but on most days of the week is 1

cause Im the best ever. No one can stop me. But overall my favorite number is 10, so yes 10." While this does give a number (10), it clearly is not a simple, numerical answer we were looking for and takes some time to interpret. All of the categorical questions had an "other" option for everyone that did not fit into the categories provided, but the quantitative questions that did not apply to all students (e.g., "What time is your curfew on weekends?" for students who do not have curfews or "What year is the car that you drive to school most often?" for students who do not drive to school) were missing "other" options. Students sometimes replied "I don't know," "other," "don't drive," "N/A," "don't have one," "when I want," etc., complicating the analysis of the results. Lastly, we should have standardized the inputs for the quantitative survey questions by specifying exactly how numbers should have been formatted, such as military time or hours. Arbitrary results such as "12" or "8" may mean either A.M. or P.M. time, and for some the sleep question, some people answered in minutes while the majority answered in hours.

Nonresponse bias

To reduce nonresponse bias, AP Statistics students were tasked with delivering their surveys to their intended recipients and follow up to make sure the survey was completed. Even though not all of the students were able to personally deliver the surveys to their recipients, the majority were able to reach the survey respondents via other students or other means.

Question	Number of Recorded Responses	Number / % of Nonresponse
What is your main writing utensil?	82	14 / 14.6%
What brand of macaroni and cheese do you prefer?	82	14 / 14.6%
On average, how many hours do you sleep per night on school nights (Sunday-Thursday)?	81	15 / 15.6%
On weekends, what time is your curfew?	82	14 / 14.6%
Where do you most frequently go for study hall?	83	13 / 13.5%
How many schools did you apply to early (early action/early decision)?	84	12 / 12.5%
What type of chocolate do you prefer?	85	11 / 11.5%
On average, how many hours on a school night (Sunday – Thursday) do you spend on homework? Studying included.	83	13 / 13.5%
What is the year of the car you drive to school most often? (students)	86	10 / 10.4%
What is the year of the car that you drive to Barlow most often? (staff)	91	54 / 37.2%

How many AP courses will you have taken by the end of high school?	83	13 / 13.5%
How many cars do the people you live with own?	83	13 / 13.5%
What is your favorite number?	82	14 / 14.6%

Of the 96 student surveys, we had an average of 13.5% nonresponse (an average of 13 missing responses), and 37.2% nonresponse for the single-question staff survey. There was also some variation between the number of responses in the student surveys (presumably human error from the data collation), ranging between 81 and 86 recorded student surveys for each question.

In other words, about 6/7 of the responses for students were collected, so that we are confident with our results because a large majority of the responses were collected. However, because the data does not indicate which individuals were not reached, it is difficult to interpret if certain parts of the student population may have been underrepresented and how that might have affected the results, and therefore we are not completely confident with our results.

The students had no control over the staff survey, which was sent out by email and prone to voluntary response bias. The high nonresponse (over one-third nonresponse) indicates that we cannot strongly trust the results of the survey.

Validity of the survey and improvements

Because we collected a large portion of the responses (with 13.5% nonresponse) for the students and carried out necessary measures to try and reduce nonresponse and response bias, we are pretty confident in the results of the student survey. There is error in the collation of the results and there may be underrepresentation of some parts of the student population, but the fact that we collected the majority of the surveys and had mostly interpretable responses means that the conclusions drawn are mostly valid.

One step to fix the some of the response bias by indicating the type of response (i.e., time to the nearest hour, numeric value, etc.) and covering all possible cases to receive more homogeneous, usable results.

For staff surveys, because of the high voluntary response bias, the same cannot be said. A better approach would have been to follow up with the survey respondents, either by repeated emailing or in person, to make sure the survey was filled out.