STAT 6800 Mid Semester Exams

Augustine Ennin

October 2025

Q1

```
DATA three;
INPUT @1 x $5. @6 y 1. @7 z 3.1; /*(a);
y2 = y**2; /*(b);
LABEL x = "xname" y = "yname"; /*(c);
DATALINES; /*(d);
One  1234
Two  5678
Three9876
Four 5432
;
RUN;
PROC PRINT DATA=three LABEL;
```

Obs	xname	yname	z	y2
1	One	1	23.4	1
2	Two	5	67.8	25
3	Three	9	87.6	81
4	Four	5	43.2	25

Q2.(a)

```
DATA src;
INFILE "/home/u63997979/sasuser.v94/Elliott and Morrell/Src.dat.txt";
INPUT ID 1-4 gender 6 env 8-9 env_qual 12 air_qual 14 health_hurt 16 pl_an_exist
    18-19 protect_env_jobs 22-23 protect_env_pop 25-26 emp_status 30-31 hrs_worked
    35-36 yrly_income 40-41 age 45-46 p_party 50-51  lib_conserv 55;
RUN;
```

Q2.(b)

```
DATA src1;
SET src;
IF nmiss(of _all_) > 0 THEN DELETE;
RUN;
PROC PRINT DATA = src1(OBS=10)
```

Obs	ID	gender	env	env_qual	air_qual	health_hurt	pl_an_exist	protect_env_jobs	protect_env_pop	emp_status	hrs_worked	yrly_income	age	p_party	lib_conserv
1	1007	5	4	3	3	5	5	5	5	1	42	20	39	1	3
2	1020	5	3	5	5	1	9	5	1	1	30	1	20	1	1
3	1023	5	5	7	7	1	9	5	5	1	50	55	42	3	3
4	1025	5	5	7	7	1	3	5	3	1	52	90	48	1	2
5	1033	1	3	5	3	1	9	7	1	1	40	15	46	1	2
6	1035	1	4	5	5	3	3	5	5	1	40	20	43	3	3
7	1039	1	8	7	7	1	9	7	7	1	30	25	29	11	3
8	1045	1	5	5	5	1	7	5	3	1	15	25	72	1	3
9	1050	5	6	5	3	1	7	7	1	1	10	25	32	1	2
10	1053	5	10	7	5	1	9	9	9	1	52	45	39	5	3

Q2.(c)

```
PROC FORMAT;
VALUE politicalfmt 1="republican" 3="democrat" 5="independent" 7="libertarian" 9="
    others" 11="no category";
RUN;

PROC PRINT DATA=src1(OBS=10);
VAR ID gender hrs_worked yrly_income p_party;
FORMAT p_party politicalfmt.;
RUN;
```

Obs	ID	gender	hrs_worked	yrly_income	p_party
1	1007	5	42	20	republican
2	1020	5	30	1	republican
3	1023	5	50	55	democrat
4	1025	5	52	90	republican
5	1033	1	40	15	republican
6	1035	1	40	20	democrat
7	1039	1	30	25	no category
8	1045	1	15	25	republican
9	1050	5	10	25	republican
10	1053	5	52	45	independent

Q2.(d)

```
PROC FORMAT;
VALUE genderfmt 1="female" 5="male";

PROC MEANS DATA=src1 MEAN STD MODE MIN Q1 MEDIAN Q3 MAX RANGE;
FORMAT gender genderfmt.;
CLASS gender;
VAR hrs_worked;
RUN;
```

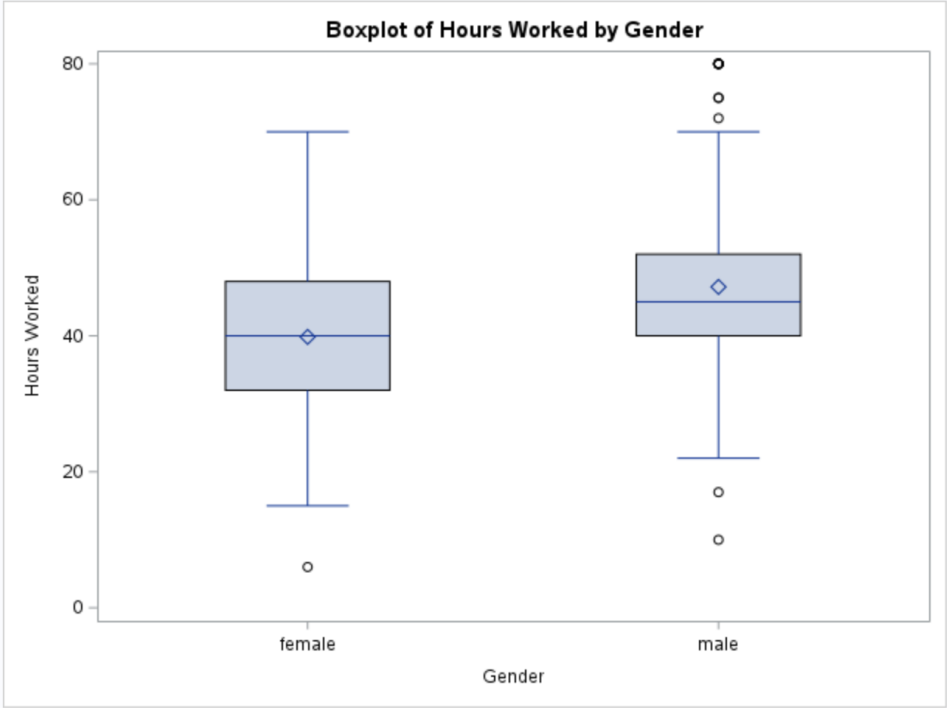
The MEANS Procedure

Analysis Variable : hrs_worked										
gender	N Obs	Mean	Std Dev	Mode	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Range
female	105	39.8380952	11.5134736	40.0000000	6.0000000	32.0000000	40.0000000	48.0000000	70.0000000	64.0000000
male	137	47.1897810	12.7074161	40.0000000	10.0000000	40.0000000	45.0000000	52.0000000	80.0000000	70.0000000

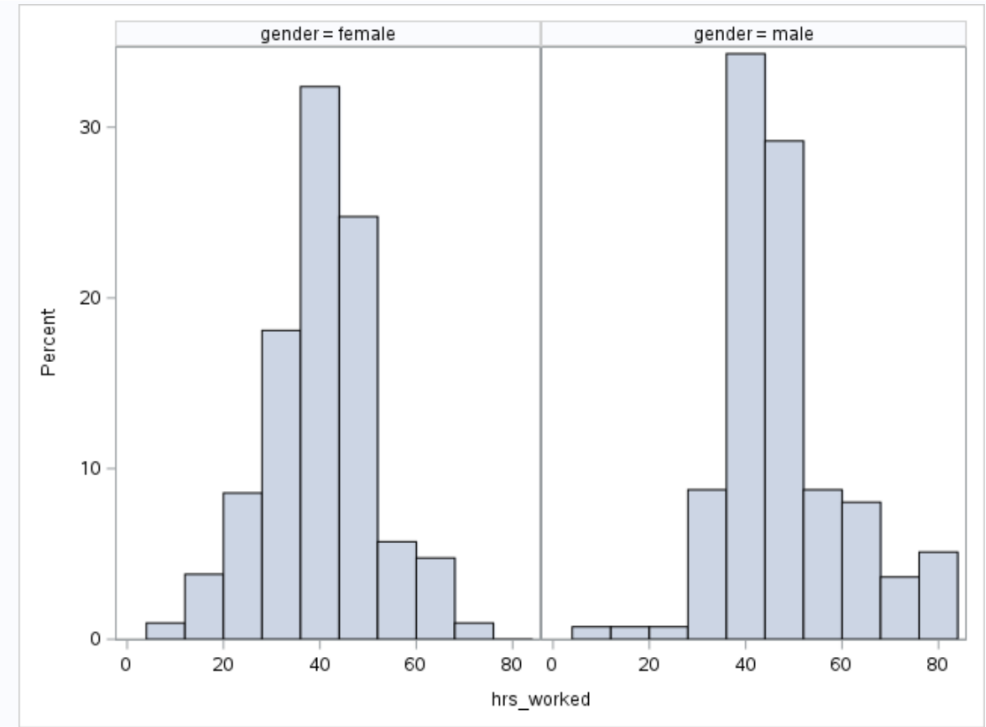
Overall, males report working longer hours than females on average, with a median of 45 vs. 40 hours. Both groups cluster around the typical 40-hour workweek, but males tend to work more overtime (up to 80 hours), whereas females show more part-time work cases (minimum as low as 6 hours).

Q2.(e)

```
PROC SGPLOT DATA = src1;
VBOX hrs_worked / CATEGORY=gender;
TITLE "Boxplot of Hours Worked by Gender";
YAXIS LABEL="Hours Worked";
XAXIS LABEL="Gender";
FORMAT gender genderfmt.;
RUN;
```




```
PROC SGPANEL DATA=src1;
PANELBY gender / LAYOUT=COLUMNLATTICE;
HISTOGRAM hrs_worked;
FORMAT gender genderfmt.;
RUN;
```



The plot shows that the median hours worked for females is slightly lower (around 40) compared to males (around 45). The spread is also tighter for females, while males show a wider range of working hours, extending to higher maximum values.

Q3

```
DATA dataset1;
INPUT EmpID Name $ Age;
DATALINES;
1 John 25
2 Mary 30
3 Alex 28
4 Sarah 35
5 Mike 40
6 Linda 27
7 Tom 33
8 Emma 29
9 Chris 31
10 Anna 26
11 James 32
12 Kate 34
13 David 45
14 Lucy 38
15 Mark 29
16 Nora 37
17 Paul 41
18 Rita 36
19 Bless 39
20 Lily 42
;
RUN;
```

```
DATA dataset2;
INPUT EmpID Salary Department $;
DATALINES;
5 70000 HR
2 55000 Finance
12 60000 IT
8 50000 Sales
1 48000 HR
16 75000 IT
3 53000 Sales
10 49000 Finance
18 72000 HR
6 51000 IT
20 80000 Finance
7 56000 Sales
11 59000 HR
9 52000 IT
14 68000 Sales
4 62000 Finance
15 54000 HR
19 73000 IT
17 71000 Sales
13 65000 Finance
;
RUN;
```

```
PROC PRINT DATA=dataset1(OBS=10);
TITLE "Dataset 1";
RUN;
```

Dataset 1

Obs	EmpID	Name	Age
1	1	John	25
2	2	Mary	30
3	3	Alex	28
4	4	Sarah	35
5	5	Mike	40
6	6	Linda	27
7	7	Tom	33
8	8	Emma	29
9	9	Chris	31
10	10	Anna	26

```
PROC PRINT DATA=dataset2(OBS=10);
TITLE "Dataset 2";
RUN;
```

Dataset 2

Obs	EmpID	Salary	Department
1	1	48000	HR
2	2	55000	Finance
3	3	53000	Sales
4	4	62000	Finance
5	5	70000	HR
6	6	51000	IT
7	7	56000	Sales
8	8	50000	Sales
9	9	52000	IT
10	10	49000	Finance

Q3.(a)

```
DATA concatdataset;
SET dataset1 dataset2;
RUN;
PROC PRINT DATA=concatdataset;
RUN;
```

Obs	EmpID	Name	Age	Salary	Department
1	1	John	25	.	
2	2	Mary	30	.	
3	3	Alex	28	.	
4	4	Sarah	35	.	
5	5	Mike	40	.	
6	6	Linda	27	.	
7	7	Tom	33	.	
8	8	Emma	29	.	
9	9	Chris	31	.	
10	10	Anna	26	.	
11	11	James	32	.	
12	12	Kate	34	.	
13	13	David	45	.	
14	14	Lucy	38	.	
15	15	Mark	29	.	
16	16	Nora	37	.	
17	17	Paul	41	.	
18	18	Rita	36	.	
19	19	Bless	39	.	
20	20	Lily	42	.	
21	1		.	48000	HR
22	2		.	55000	Finance
23	3		.	53000	Sales
24	4		.	62000	Finance
25	5		.	70000	HR
26	6		.	51000	IT
27	7		.	56000	Sales
28	8		.	50000	Sales
29	9		.	52000	IT
30	10		.	49000	Finance
31	11		.	59000	HR
32	12		.	60000	IT
33	13		.	65000	Finance
34	14		.	68000	Sales
35	15		.	54000	HR
36	16		.	75000	IT
37	17		.	71000	Sales
38	18		.	72000	HR
39	19		.	73000	IT
40	20		.	80000	Finance

Q3.(b)

```
PROC SORT DATA=dataset1;
BY EmpID;
RUN;

PROC SORT DATA=dataset2;
BY EmpID;
RUN;

DATA interleavedataset;
SET dataset1 dataset2;
BY EmpID;
RUN;

PROC PRINT DATA=interleavedataset;
RUN;
```

Obs	EmpID	Name	Age	Salary	Department
1	1	John	25	.	
2	1		.	48000	HR
3	2	Mary	30	.	
4	2		.	55000	Finance
5	3	Alex	28	.	
6	3		.	53000	Sales
7	4	Sarah	35	.	
8	4		.	62000	Finance
9	5	Mike	40	.	
10	5		.	70000	HR
11	6	Linda	27	.	
12	6		.	51000	IT
13	7	Tom	33	.	
14	7		.	56000	Sales
15	8	Emma	29	.	
16	8		.	50000	Sales
17	9	Chris	31	.	
18	9		.	52000	IT
19	10	Anna	26	.	
20	10		.	49000	Finance
21	11	James	32	.	
22	11		.	59000	HR
23	12	Kate	34	.	
24	12		.	60000	IT
25	13	David	45	.	
26	13		.	65000	Finance
27	14	Lucy	38	.	
28	14		.	68000	Sales
29	15	Mark	29	.	
30	15		.	54000	HR
31	16	Nora	37	.	
32	16		.	75000	IT
33	17	Paul	41	.	
34	17		.	71000	Sales
35	18	Rita	36	.	
36	18		.	72000	HR
37	19	Bless	39	.	
38	19		.	73000	IT
39	20	Lily	42	.	
40	20		.	80000	Finance

Q3.(c)

```
PROC SORT DATA=dataset1;
BY EmpID;
RUN;

PROC SORT DATA=dataset2;
BY EmpID;
RUN;

DATA mergeddataset;
MERGE dataset1 dataset2;
BY EmpID;
RUN;
PROC PRINT DATA=mergeddataset;
RUN;
```

Obs	EmpID	Name	Age	Salary	Department
1	1	John	25	48000	HR
2	2	Mary	30	55000	Finance
3	3	Alex	28	53000	Sales
4	4	Sarah	35	62000	Finance
5	5	Mike	40	70000	HR
6	6	Linda	27	51000	IT
7	7	Tom	33	56000	Sales
8	8	Emma	29	50000	Sales
9	9	Chris	31	52000	IT
10	10	Anna	26	49000	Finance
11	11	James	32	59000	HR
12	12	Kate	34	60000	IT
13	13	David	45	65000	Finance
14	14	Lucy	38	68000	Sales
15	15	Mark	29	54000	HR
16	16	Nora	37	75000	IT
17	17	Paul	41	71000	Sales
18	18	Rita	36	72000	HR
19	19	Bless	39	73000	IT
20	20	Lily	42	80000	Finance

Q4.

```
DATA scleroderma;
INFILE "/home/u63997979/sasuser.v94/Elliott and Morrell/Sclero.dat";
INPUT clinic_no 1-2 ID 4-5 treatment 8 skin_thick1 11-12 skin_thick2 15-16
    skin_mobility1 19-21 skin_mobility2 24-26 patient_assess1 29 patient_assess2
    32;

```

```
/*Improvement Indicators;
IF skin_thick1 > skin_thick2 THEN thick_improve = 1; ELSE thick_improve = 0;
IF skin_mobility1 < skin_mobility2 THEN mob_improve = 1; ELSE mob_improve = 0;
IF patient_assess1 > patient_assess2 THEN assess_improve = 1; ELSE assess_improve
    = 0;

/*Labels for the indicator variables;
LABEL thick_improve = "Improved if skin thickening decreased"
          mob_improve = "Improved if skin mobility increased"
          assess_improve = "Improved if patient assessment decreased";
RUN;

PROC PRINT DATA=scleroderma(OBS=10) LABEL;
TITLE "First 10 Observation of the Sclero Dataset";
RUN;
```

First 10 Observation of the Sclero Dataset

Obs	clinic_no	ID	treatment	skin_thick1	skin_thick2	skin_mobility1	skin_mobility2	patient_assess1	patient_assess2	Improved if skin thickening decreased	Improved if skin mobility increased	Improved if patient assessment decreased
1	1	1	2	.	.	298	418	8	5	0	1	1
2	1	2	1	.	.	390	498	3	3	0	1	0
3	1	3	2	.	.	543	658	2	2	0	1	0
4	1	4	1	.	.	526	535	2	3	0	1	0
5	2	5	1	30	28	227	260	6	2	1	1	1
6	2	6	2	18	21	375	409	6	5	0	1	1
7	8	7	2	27	25	308	306	6	3	1	0	1
8	8	8	2	19	25	376	360	2	3	0	0	0
9	8	9	1	18	19	281	377	3	3	0	1	0
10	8	10	1	25	21	264	314	4	3	1	1	1

```
PROC FREQ DATA=scleroderma;
TABLES clinic_no*treatment;
```

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of clinic_no by treatment

clinic_no	treatment 1	treatment 2	Total
1	2 2.63 50.00 6.25	2 2.63 50.00 4.55	4 5.26
2	1 1.32 50.00 3.13	1 1.32 50.00 2.27	2 2.63
8	2 2.63 50.00 6.25	2 2.63 50.00 4.55	4 5.26
10	4 5.26 57.14 12.50	3 3.95 42.86 6.82	7 9.21
18	0 0.00 0.00 0.00	4 5.26 100.00 9.09	4 5.26
40	1 1.32 25.00 3.13	3 3.95 75.00 6.82	4 5.26
41	4 5.26 66.67 12.50	2 2.63 33.33 4.55	6 7.89
42	0 0.00 0.00 0.00	1 1.32 100.00 2.27	1 1.32
45	4 5.26 57.14 12.50	3 3.95 42.86 6.82	7 9.21
46	4 5.26 36.36 12.50	7 9.21 63.64 15.91	11 14.47
47	1 1.32 50.00 3.13	1 1.32 50.00 2.27	2 2.63
48	3 3.95 30.00 9.38	7 9.21 70.00 15.91	10 13.16
49	5 6.58 45.45 15.63	6 7.89 54.55 13.64	11 14.47
50	1 1.32 33.33 3.13	2 2.63 66.67 4.55	3 3.95
Total	32 42.11	44 57.89	76 100.00

Q4.(a) Clinics number 46 and 49 had 11 patients each, which is the largest among all the clinics.

Q4.(b)

Table 1: Clinic Row Percentages

Clinic Number	**Percentage**
Clinic 1	50%
Clinic 2	50%
Clinic 8	50%
Clinic 10	42.86%
Clinic 18	100%
Clinic 40	75%
Clinic 41	33.33%
Clinic 42	100%
Clinic 45	42.86%
Clinic 46	63.64%
Clinic 47	50%
Clinic 48	70%
Clinic 49	54.55%
Clinic 50	63.67%

The percent of subjects in the control group (placebo=2) ranges from 33.33% (Clinic 41) to 100% (Clinics 18 and 42) depending on the clinic.

Q4.(c)

```
PROC FREQ DATA=scleroderma;
WHERE clinic_no in (46, 48, 49);
TABLES treatment*thick_improve
           treatment*mob_improve
           treatment*assess_improve / NOCOL NOPERCENT;
TITLE "Assessing Patients' Conditions for Clinics 46, 48, and 49";
RUN;
```

Assessing Patients' Conditions for Clinics 46, 48, and 49

The FREQ Procedure

Frequency
Row Pct

Table of treatment by thick_improve			
	thick_improve(Improved if skin thickening decreased)		
treatment	0	1	Total
1	8 66.67	4 33.33	12
2	17 85.00	3 15.00	20
Total	25	7	32

Frequency
Row Pct

Table of treatment by mob_improve			
	mob_improve(Improved if skin mobility increased)		
treatment	0	1	Total
1	3 25.00	9 75.00	12
2	12 60.00	8 40.00	20
Total	15	17	32

Frequency
Row Pct

Table of treatment by assess_improve			
	assess_improve(Improved if patient assessment decreased)		
treatment	0	1	Total
1	8 66.67	4 33.33	12
2	14 70.00	6 30.00	20
Total	22	10	32

Across Clinics 46, 48, and 49, the rates of patient improvement varied by treatment type. In thickening of the skin, 33.3% of patients improved under treatment 1 (drug) compared to just 15.0% under treatment 2 (placebo). In mobility of the skin, 75.0% of patients in treatment 1 improved compared to 40.0% in treatment 2. For assessment improvement, the rates were similar under the two treatments (33.3% for treatment 1 and 30.0% for treatment 2).Overall, treatment 1 (drug) showed systematically higher rates of improvement in all three clinics.

Q4.(d)

```
PROC FREQ DATA=scleroderma;
WHERE treatment=1;
TABLES thick_improve*assess_improve;
TITLE "Skin Thickening  vs. Patient Assessment for the Drug Group";
RUN;
```

Skin Thickening vs. Patient Assessment for the Drug Group

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of thick_improve by assess_improve			
thick_improve(Improved if skin thickening decreased)	assess_improve(Improved if patient assessment decreased) 0	1	Total
0	11 34.38 64.71 57.89	6 18.75 35.29 46.15	17 53.13
1	8 25.00 53.33 42.11	7 21.88 46.67 53.85	15 46.88
Total	19 59.38	13 40.63	32 100.00

34.38% had bad outcomes in both measurement while 21.88% had good outcomes in both measurement.

Q5

```
PROC IMPORT DATAFILE="/home/u63997979/sasuser.v94/Elliott and Morrell/studdata043.
    xls"
     OUT=studdata
     DBMS=xls
     REPLACE;
     SHEET="Sheet1";
     GETNAMES=YES;
RUN;

PROC PRINT DATA=studdata (OBS=10);
TITLE "First 10 Observation of the studdata";
RUN;
```

First 10 Observation of the studdata

Obs	No.	Gender	Age	Class	GPA	Height	Weight	BMI	WtGrp	Hand	Drinks	Exercise	Smoker	SmokFree	Credits	HW	Q
1	1	M	20	Sophomore	3.5	71	185	25.85449	Overwt	R	0	1.5	N	Y	15	2	
2	2	M	18	Junior	3.55	74	186	23.92933	Normal	L	0	7	N	Y	21	18	
3	3	F	20	Sophomore	4	69	155	22.93583	Normal	R	30	3	N	N	15	10	
4	4	F	19	Sophomore	3.56	65	135	22.51065	Normal	L	5	2	N	Y	17	30	
5	5	F	21	Junior	3.4	67	135	21.18679	Normal	R	35	3	N	Y	13	10	
6	6	F	23	Gradduate	3.7	64	130	22.35962	Normal	R	0	1	N	Y	11	12	
7	7	M	23	Senior	3.1	69	165	24.41556	Normal	R	2	5	N	Y	19	8	
8	8	M	19	Sophomore	4	72	215	29.21827	Overwt	L	2	1.5	N	N	18	8	
9	9	M	21	Senior	3.4	75	175	21.91778	Normal	R	8	5	N	Y	18	9	
10	10	M	23	Senior	2.85	74.5	210	26.65556	Overwt	R	12	15	N	Y	17	12.5	

Q5.(a)

$H_0 : \mu = 3.3$ vs. $H_1 : \mu \neq 3.3$

```
PROC TTEST DATA=studdata H0=3.3;
```

```
VAR GPA;
TITLE "Test of Mean GPA Against Historical Average of 3.3";
RUN;
```

Test of Mean GPA Against Historical Average of 3.3

The TTEST Procedure

Variable: GPA (GPA)

N	Mean	Std Dev	Std Err	Minimum	Maximum
131	3.1521	0.5284	0.0462	1.7000	4.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3.1521	3.0607	3.2434	0.5284	0.4712	0.6015

DF	t Value	Pr > |t|
130	-3.20	0.0017

Since the $p-value$ is 0.0017 (< 0.05), we reject the null hypothesis. There is strong evidence that the mean GPA differs from the historical average of 3.3.

Q5.(b)

H_0 : Male mean = Female mean (males and females spend the same time exercising)
H_1 : Male mean > Female mean (males spend more time exercising)

```
PROC TTEST DATA=studdata SIDES=U;
CLASS Gender;
VAR Exercise;
TITLE "Test If Males Exercise More than Females";
RUN;
```

Test If Males Exercise More than Females

The TTEST Procedure

Variable: Exercise (Exercise)

Gender	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
F		64	3.5938	4.3715	0.5464	0	20.0000
M		67	3.6642	3.3171	0.4052	0	15.0000
Diff (1-2)	Pooled		-0.0704	3.8681	0.6761		
Diff (1-2)	Satterthwaite		-0.0704		0.6803		

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
F		3.5938	2.5018	4.6857	4.3715	3.7236	5.2944
M		3.6642	2.8551	4.4733	3.3171	2.8351	3.9981
Diff (1-2)	Pooled	-0.0704	-1.1905	Infty	3.8681	3.4482	4.4054
Diff (1-2)	Satterthwaite	-0.0704	-1.1983	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	129	-0.10	0.5414
Satterthwaite	Unequal	117.44	-0.10	0.5411

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	63	66	1.74	0.0278

The equality of variances test was significant (p = 0.0278), so unequal variances were assumed. The two-sample t-test (Satterthwaite) gave t = -0.10 with p = 0.54. Since the p-value is much greater than 0.05, we fail to reject the null hypothesis. There is no evidence that male students spend more hours exercising than female students.

Q5.(c)

H_0 : Male proportion = Female proportion (males and females have the same proportion of smokers)
H_1 : Male proportion $\neq$ Female proportion (different proportion of male and female smokers)

```
PROC FREQ DATA = studdata;
TABLES Gender*Smoker / CHISQ;
TITLE "Test for Proportion of Males and Females Smokers";
RUN;
```

Test for Proportion of Males and Females Smokers

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Gender by Smoker		
	Smoker(Smoker)		
Gender(Gender)	N	Y	Total
F	49 37.40 76.56 45.37	15 11.45 23.44 65.22	64 48.85
M	59 45.04 88.06 54.63	8 6.11 11.94 34.78	67 51.15
Total	108 82.44	23 17.56	131 100.00

Statistics for Table of Gender by Smoker

Statistic	DF	Value	Prob
Chi-Square	1	2.9892	0.0838
Likelihood Ratio Chi-Square	1	3.0232	0.0821
Continuity Adj. Chi-Square	1	2.2477	0.1338
Mantel-Haenszel Chi-Square	1	2.9664	0.0850
Phi Coefficient		-0.1511	
Contingency Coefficient		0.1494	
Cramer's V		-0.1511	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	49
Left-sided Pr <= F	0.0665
Right-sided Pr >= F	0.9755
Table Probability (P)	0.0421
Two-sided Pr <= P	0.1086

Sample Size = 131

Based on the chi-square test ($\chi^2 = 2.99,\ p = 0.084$), there is no evidence of a significant difference in smoking proportion between male and female students.