STAT4950
Fall, 2006

HOMEWORK/LAB 1

Question 1:

Note that the sample size on the homework sheet is 21, while that provided in lab is 25. The results are similar.

1. Refer to Table 1
    a. The percent of men is 52%.
    b. The mode for marital status is 2 (Divorced).
    c. The frequency of divorced people in the sample is 11

Table 1

FREQUENCY TABLE OF GENDER AND MARITAL STATUS 1
19:24 Wednesday, October 5, 2005

The FREQ Procedure

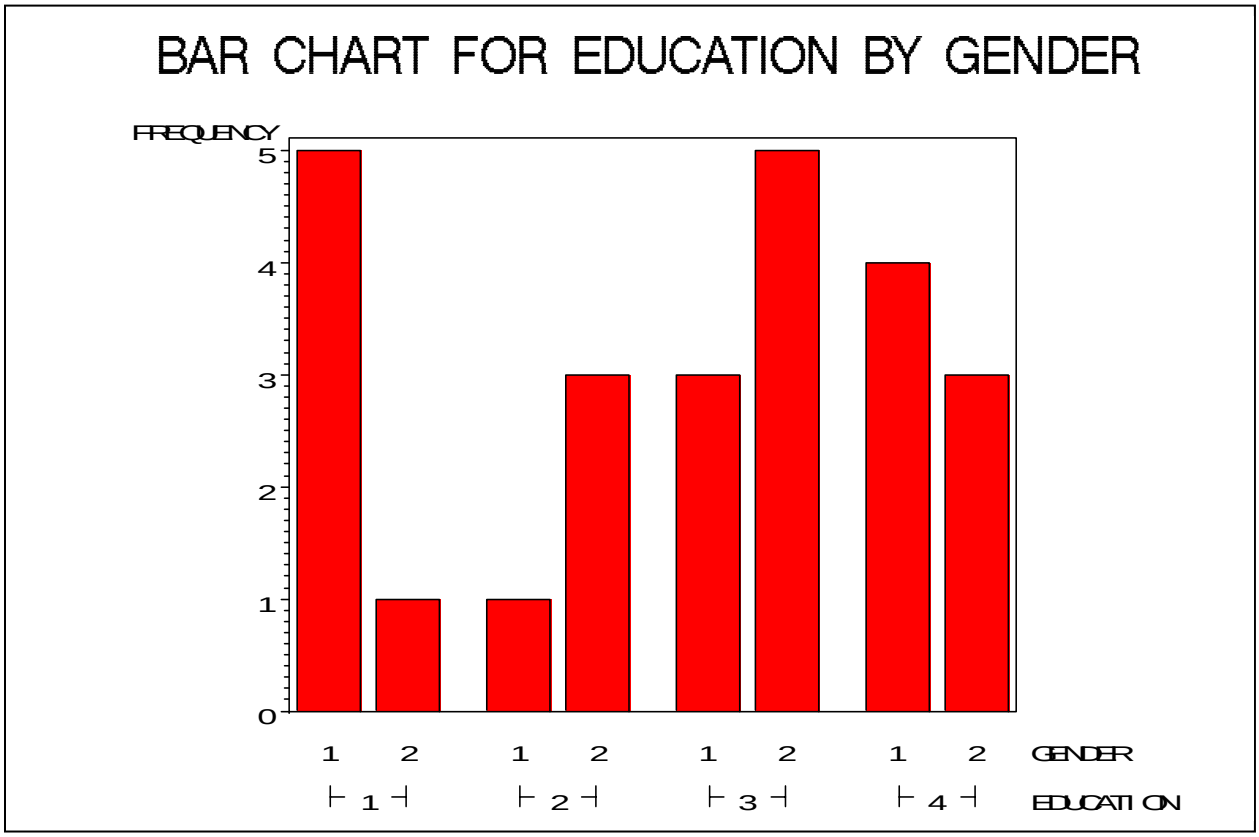
GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	13	52.00	13	52.00
2	12	48.00	25	100.00

MARITAL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	9	36.00	9	36.00
2	11	44.00	20	80.00
3	5	20.00	25	100.00

2. Refer to Figure 1

According to the bar chart, there are more men with below high school and postgraduate education, and in contrast there are more women at the high school graduates and college graduate levels.

Figure 1



3. Refer to Table 2

Table 2

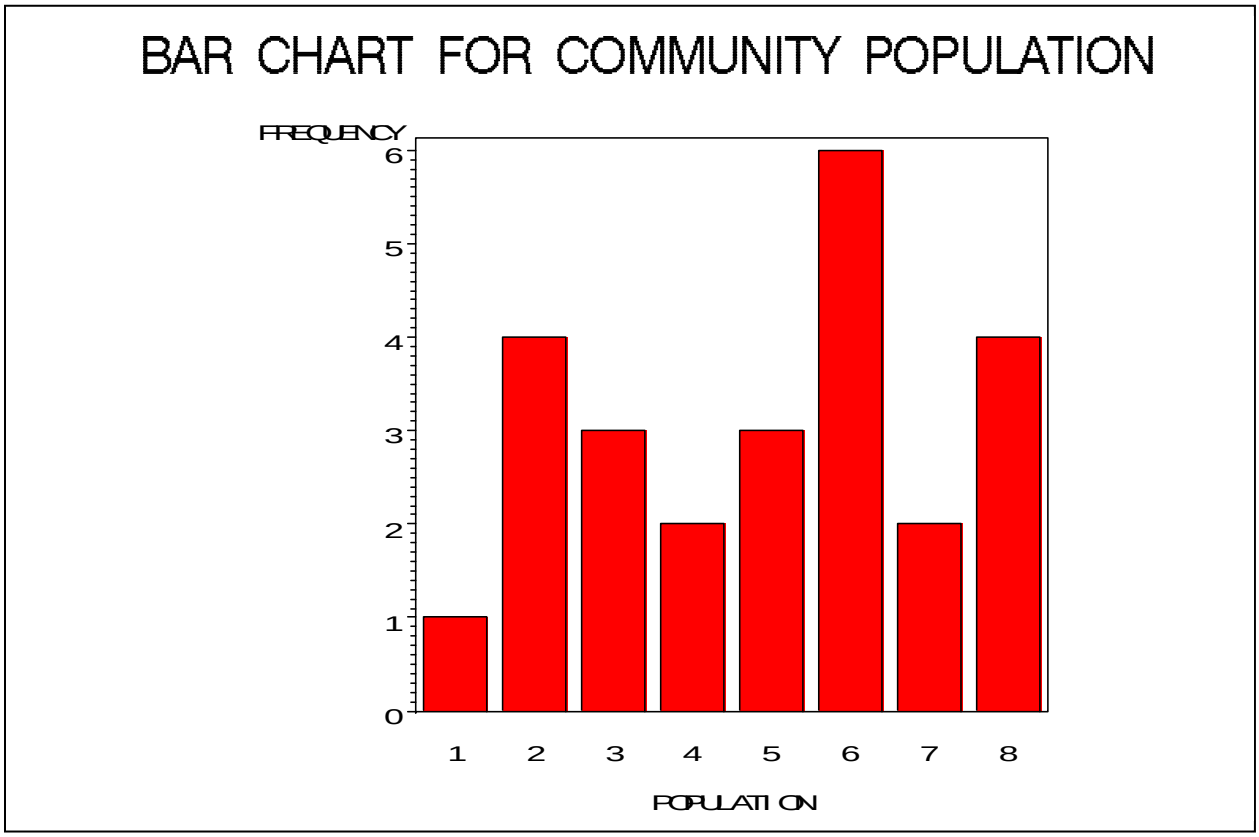
FREQUENCY TABLE FOR EDUCATION 2
19:24 Wednesday, October 5, 2005

The FREQ Procedure

EDUCATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	24.00	6	24.00
2	4	16.00	10	40.00
3	8	32.00	18	72.00
4	7	28.00	25	100.00

4. Refer to Figure 2

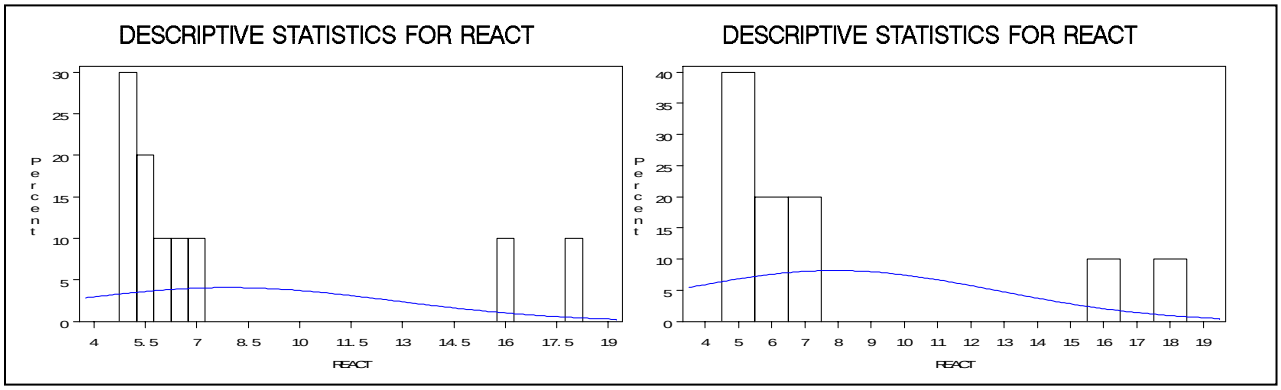
Figure 2



5. There are about the same number of men as women. In the Ann sample, divorced people are more prevalent and never-married people are less prevalent. A large number of men did not graduate from high school while a large number of women are college graduates. Most people live in communities with a population between 1,001 to 5,000 people.

Question 2:

1. Two histograms for REACT:

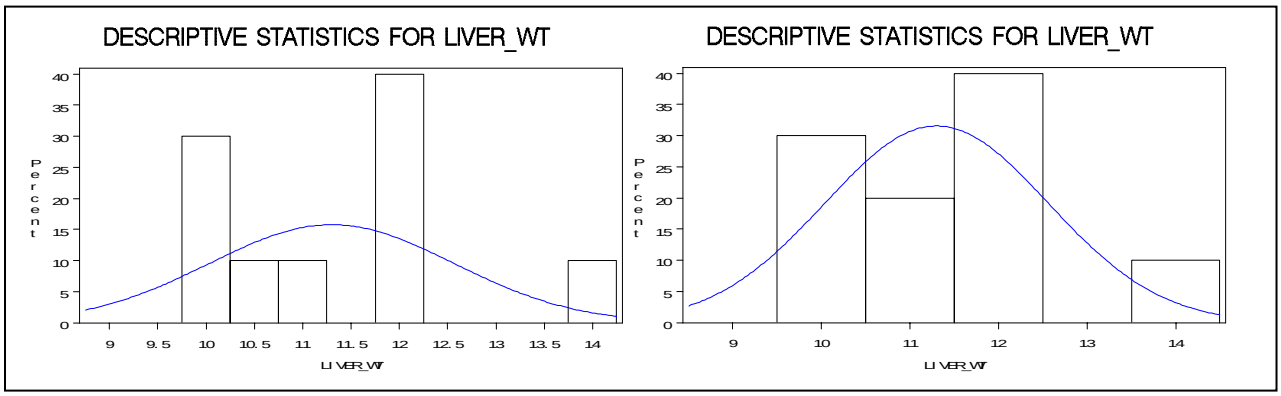


The SAS output for boxplot, normality tests and normal probability plot are shown below. Since the p-values are all small and the points on the plot vary from the line, the assumption that REACT is normally distributed is failed.

```
REACT                              Tests for Normality

                Test                  --Statistic---    -----p Value------

                Shapiro-Wilk          W     0.656993    Pr < W      0.0003
                Kolmogorov-Smirnov    D      0.38205    Pr > D     <0.0100
                Cramer-von Mises      W-Sq  0.315825    Pr > W-Sq  <0.0050
                Anderson-Darling      A-Sq  1.628355    Pr > A-Sq  <0.0050

                    Stem Leaf                     #             Boxplot
                       18 2                       1                *
                       16                                          |
                       14 8                       1                *
                       12                                          |
                       10                                          |
                        8                                          |
                        6 79                      2               +--+--+
                        4 890459                  6               *-----*
                          ----+----+----+----+

                                    Normal Probability Plot
                    19+                                        *     ++++
                      |                                      ++++
                      |                                  *  ++++
                      |                                ++++
                      |                            ++++
                      |                          ++++
                      |                       ++++   *  *
                     5+          *    * +*+++*  * *
                      +----+----+----+----+----+----+----+----+----+----+
                           -2        -1         0        +1        +2
```

Two

histograms for LIVER_WT:

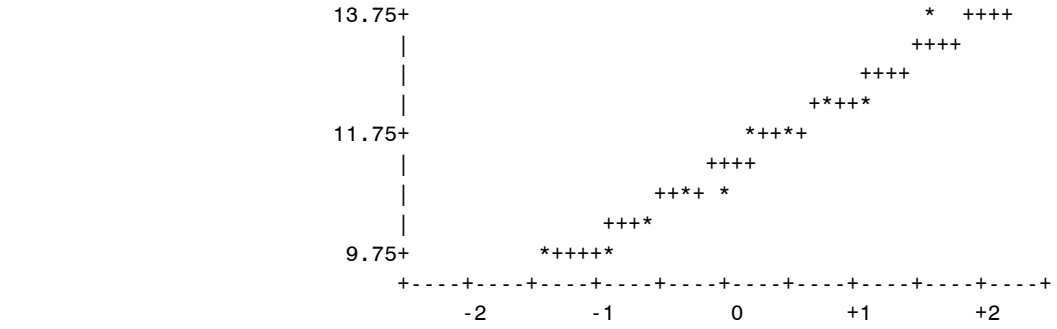


LIVER_WT — Tests for Normality

Test		--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.924078	Pr < W	0.3922
Kolmogorov-Smirnov	D	0.153785	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.051401	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.341284	Pr > A-Sq	>0.2500

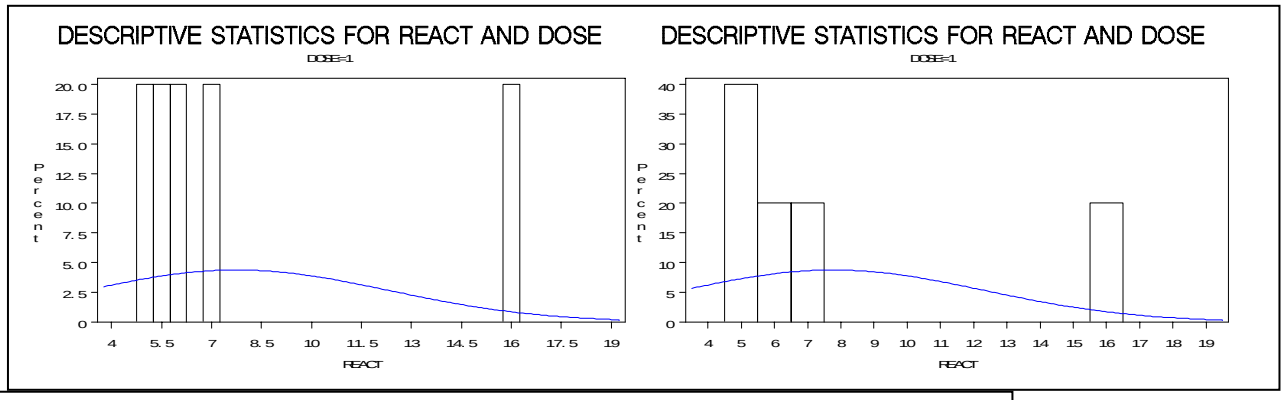
Stem	Leaf	#	Boxplot
13	8	1	
13			
12			
12	02	2	
11	89	2	
11			
10	59	2	
10	2	1	
9	89	2	

Normal Probability Plot



The SAS output for boxplot, normality tests and normal probability plot are shown above. Since the p-values are not small and the points on the plot are closed to the line, the assumption that LIVER_WT is normally distributed is reasonable.

Two histograms of REACT at Dose 1:



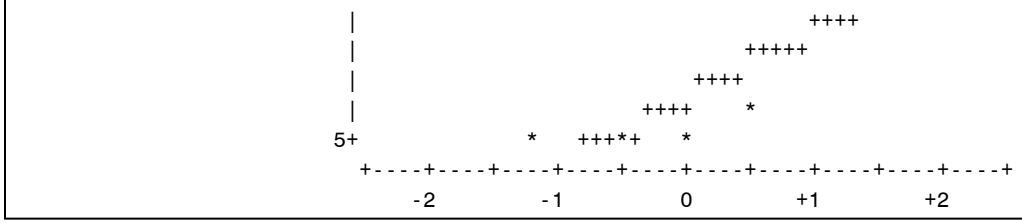
REACT AND DOSE 1 — Tests for Normality

Test		--Statistic---		-----p Value------
Shapiro-Wilk	W	0.707883	Pr < W	0.0116
Kolmogorov-Smirnov	D	0.374797	Pr > D	0.0209
Cramer-von Mises	W-Sq	0.145986	Pr > W-Sq	0.0187
Anderson-Darling	A-Sq	0.772171	Pr > A-Sq	0.0172

Stem	Leaf	#	Boxplot
14	8	1	*
12			
10			
8			
6	9	1	+--+--+
4	849	3	*-----*

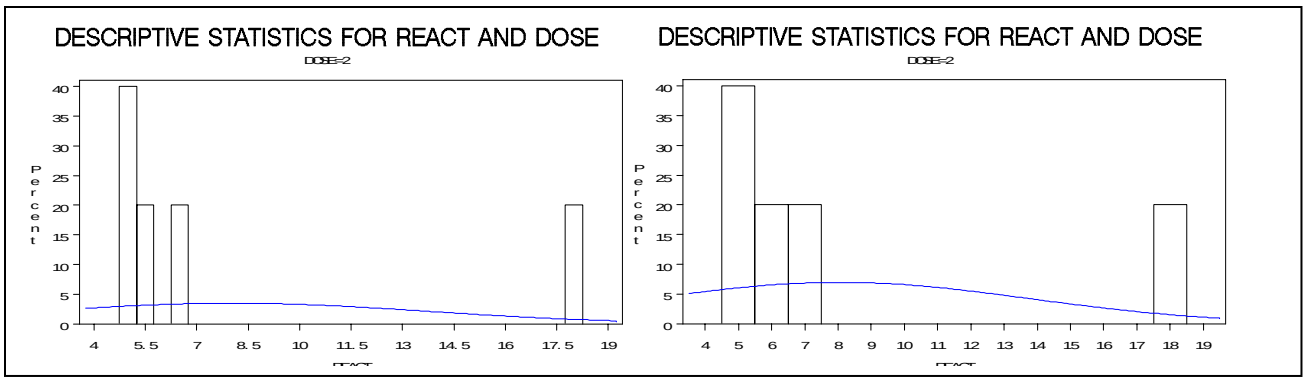
----+----+----+----+

Normal Probability Plot



The SAS output for boxplot, normality tests and normal probability plot are shown above. Since the p-values are small and the points on the plot vary from the line, the assumption that REACT at Dose 1 is normally distributed is failed.

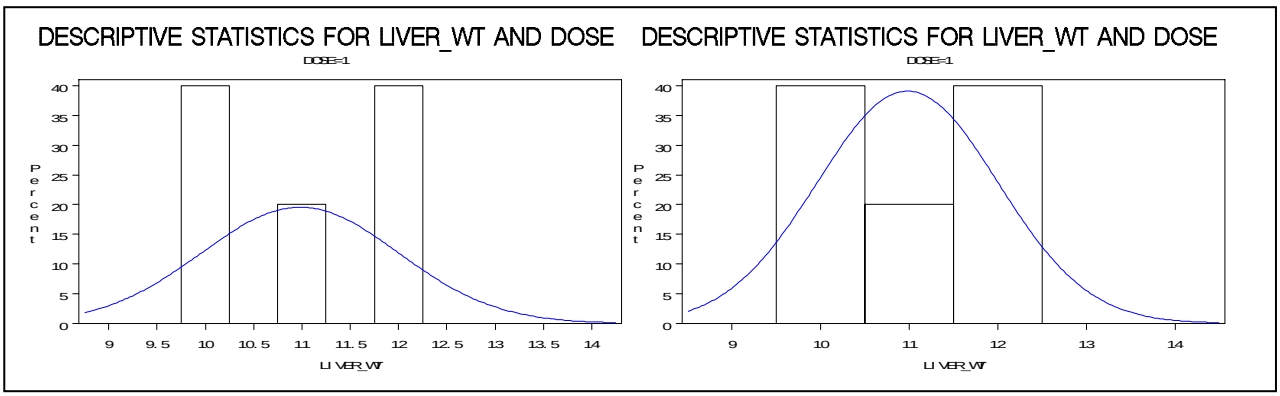
Two histograms for REACT at Dose 2:




```
REACT AND DOSE 2                    Tests for Normality

                Test                  --Statistic---    -----p Value------

                Shapiro-Wilk          W     0.655292    Pr < W       0.0031
                Kolmogorov-Smirnov    D     0.394074    Pr > D       0.0115
                Cramer-von Mises      W-Sq  0.173059    Pr > W-Sq    0.0073
                Anderson-Darling      A-Sq  0.898316    Pr > A-Sq    0.0068

                    Stem Leaf                     #             Boxplot
                       18 2                       1                *
                       16
                       14
                       12
                       10
                        8
                        6 7                       1             +-----+
                        4 905                     3             *--+--*
                          ----+----+----+----+

Normal Probability Plot
                     19+                                      *     +++
                       |                                    ++++
                       |                                  +++
                       |                                ++++
                       |                              +++
                       |                            ++++
                       |                          +++     *
                      5+              *    ++*+    *
                        +----+----+----+----+----+----+----+----+----+----+
                            -2         -1          0         +1         +2
```

The SAS output for boxplot, normality tests and normal probability plot are shown above. Since the p-values are small and the points on the plot vary from the line, the assumption that REACT at Dose 2 is normally distributed is failed.

Two histograms for LIVER_WT at Dose 1:

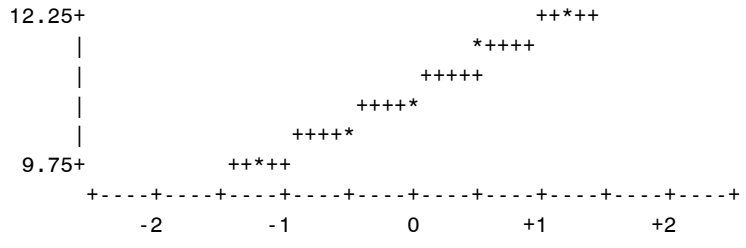


LIVER_WT AND DOSE 1 — Tests for Normality

Test		--Statistic---	-----p Value------	
Shapiro-Wilk	W	0.94184	Pr < W	0.6790
Kolmogorov-Smirnov	D	0.189101	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.032428	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.218264	Pr > A-Sq	>0.2500

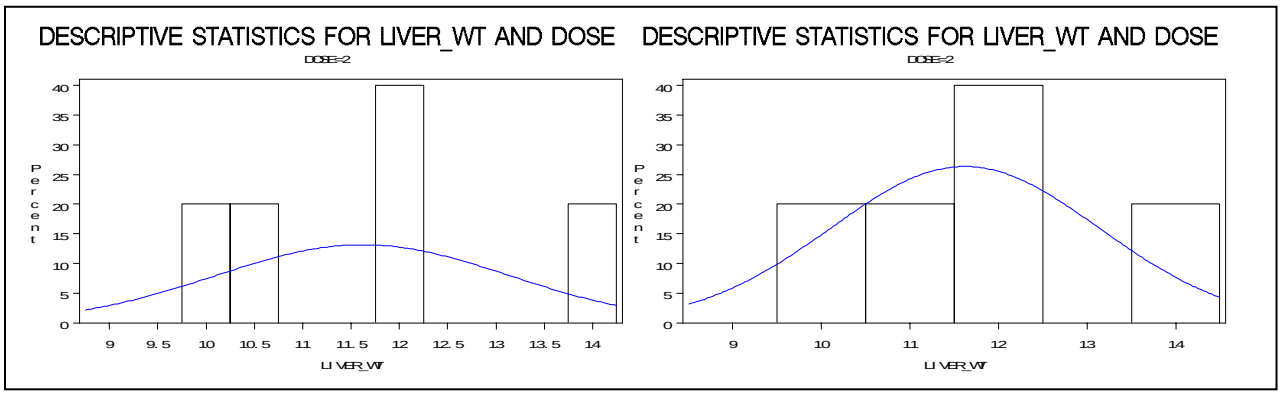
Stem	Leaf	#	Boxplot
12	2	1	
11	8	1	
11			
10	9	1	
10	2	1	
9	8	1	

Normal Probability Plot



The SAS output for boxplot, normality tests and normal probability plot are shown above. Since the p-values are not small and the points on the plot are close to the line, the assumption that LIVER_WT at Dose 1 is normally distributed is reasonable.

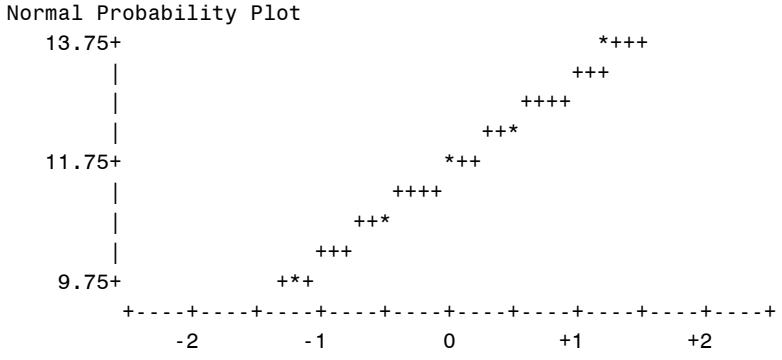
Two histograms for LIVER_WT at Dose 2:



LIVER_WT AND DOSE 2 — Tests for Normality

Test		--Statistic---		-----p Value------
Shapiro-Wilk	W	0.949815	Pr < W	0.7359
Kolmogorov-Smirnov	D	0.201012	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.03856	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.237599	Pr > A-Sq	>0.2500

Stem	Leaf	#	Boxplot
13	8	1	
13			
12			
12	0	1	
11	9	1	
11			
10	5	1	
10			
9	9	1	



The SAS output for boxplot, normality tests and normal probability plot are shown above. Since the p-values are not small and the points on the plot are close to the line, the assumption that LIVER_WT at Dose 2 is normally distributed is reasonable.

3. Because the normality assumption for REACT is failed, t-test will not be appropriate. Here is the result of a non-parametric test:

Kruskal-Wallis Test

Chi-Square	0.0109
DF	1
Pr > Chi-Square	0.9168

The p-value 0.92 is too large to reject the null hypothesis. Therefore, we have insufficient evidence to claim there is a significant difference between REACT for DOSE 1 and REACT for DOSE 2.
(Note that you will get full credit if you perform a t-test since this non-parametric test has not been taught yet.)

4.

The TTEST Procedure

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
LIVER_WT	Pooled	Equal	8	-0.78	0.4561
LIVER_WT	Satterthwaite	Unequal	7.01	-0.78	0.4592

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
LIVER_WT	Folded F	4	4	2.20	0.4628

The p-value 0.46 of the test of the equality of variances indicates that it is reasonable to assume equal variances. In addition, from the previous parts, the normal assumption for LIVER_WT is also reasonable. Therefore, we can use (pooled) t-test. The p-value is 0.456 and so we cannot reject the null hypothesis. In conclusion, we have insufficient evidence to claim there is a significant difference between LIVER_WT for DOSE 1 and LIVER_WT for DOSE 2.

Question 3: page 66, #2.9

2nd and 3rd lines: would have looked better with INPUT AGE
STATUS ($ missing)
PROGNOSIS ($ missing)
DOCTOR ($ missing)
GENDER ($ missing)
STATUS2
STATUS3;

4th line: should read DATALINES;
(data lines)
;

6th and 7th lines: should read PROC GCHART DATA=123; (the data name must start at a letter)
BY GENDER; (should have been sorted first – PROC SORT)
VBAR STATUS (missing ;)

11th and 12th lines: should be written as
PLOT DOCTOR*PROGNOSIS;

Optional Question: How can we draw a pie chart using SAS?

```
PROC GCHART DATA=______;
      TITLE "PIE CHART";
      PIE ______ / DISCRETE
                   VALUE=INSIDE
                   PERCENT=INSIDE
                   SLICE=OUTSIDE;
RUN;
```

Use the **discrete** option to insure that only the values in the dataset label slices in the pie chart.
Other options include: **value=inside** causes the frequency count to be placed inside the pie slice.
percent=inside causes the percent to be placed inside the pie slice.
slice=outside causes the label to be placed outside the pie slice.

SAS Code for STAT4950 Homework #1

```
***PROGRAM TO CREATE AND DESCRIBE DEMOGRAPHIC CHARACTERISTICS OF "ANNDATA";

DATA ANNDATA;
       INPUT   SUBJECT $
                      GENDER $
                      EDUCATION $
                      MARITAL $
                      POPULATION $;
DATALINES;
1 2 4 2 2
2 1 4 3 2
3 1 3 1 2
4 1 1 2 8
5 2 2 2 3
6 1 1 1 6
7 1 3 2 6
8 1 4 3 5
9 1 1 2 8
10 2 2 1 7
11 2 3 2 8
12 2 4 2 5
13 1 1 2 7
14 1 3 3 6
15 2 1 1 4
16 1 4 1 4
17 2 3 2 1
18 1 2 2 2
19 2 3 2 5
20 2 3 1 8
21 2 4 3 6
22 1 1 1 6
23 2 2 1 3
24 2 3 1 3
25 1 4 3 6
;

**FREQUENCY ANALYSIS OF GENDER AND MARITAL STATUS VARIABLES;

PROC FREQ DATA=ANNDATA;
       TITLE "FREQUENCY TABLE OF GENDER AND MARITAL STATUS" ;
       TABLES GENDER MARITAL;
RUN;

**GRAPHICAL SUMMARY FOR GENDER AND EDUCATION VARIABLES;

PROC GCHART DATA=ANNDATA;
       TITLE "BAR CHART FOR EDUCATION BY GENDER";
       VBAR GENDER / GROUP=EDUCATION;
RUN;

**FREQUENCY TABLE FOR EDUCATION LEVEL;
```

```
PROC FREQ DATA=ANNDATA;
        TITLE "FREQUENCY TABLE FOR EDUCATION";
        TABLES EDUCATION;
RUN;

**GRAPHICAL SUMMARY FOR COMMUNITY POPULATION;

PROC GCHART DATA=ANNDATA;
        TITLE "BAR CHART FOR COMMUNITY POPULATION";
        VBAR POPULATION;
RUN;

*************************************************************************************************
***PROGRAM TO DESCRIBE RESULTS OF EXPERIMENT FOR DATA CALLED "LIVEREXP";

DATA LIVEREXP;
        INPUT   SUBJECT $
                        DOSE $
                        REACT
                        LIVER_WT;
DATALINES;
1 1 5.4 10.2
2 1 5.9 9.8
3 1 4.8 12.2
4 1 6.9 11.8
5 1 15.8 10.9
6 2 4.9 13.8
7 2 5.0 12.0
8 2 6.7 10.5
9 2 18.2 11.9
10 2 5.5 9.9
;

**NUMERICAL AND GRAPHICAL SUMMARIES FOR ALL VARIABLES;

PROC UNIVARIATE DATA=LIVEREXP NORMAL PLOT;
        TITLE "DESCRIPTIVE STATISTICS FOR REACT";
        VAR REACT;
        HISTOGRAM REACT / MIDPOINTS=4.0 TO 19.0 BY 0.5 NORMAL;
        HISTOGRAM REACT / MIDPOINTS=4.0 TO 19.0 BY 1.0 NORMAL;
RUN;

PROC UNIVARIATE DATA=LIVEREXP NORMAL PLOT;
        TITLE "DESCRIPTIVE STATISTICS FOR LIVER_WT";
        VAR LIVER_WT;
        HISTOGRAM LIVER_WT / MIDPOINTS=9.0 TO 14.0 BY 0.5 NORMAL;
        HISTOGRAM LIVER_WT / MIDPOINTS=9.0 TO 14.0 BY 1.0 NORMAL;
RUN;

PROC SORT DATA=LIVEREXP;
        BY DOSE;
RUN;

PROC UNIVARIATE DATA=LIVEREXP NORMAL PLOT;
        TITLE "DESCRIPTIVE STATISTICS FOR REACT AND DOSE";
        BY DOSE;
```

```
        VAR REACT;
        HISTOGRAM REACT / MIDPOINTS=4.0 TO 19.0 BY 0.5 NORMAL;
        HISTOGRAM REACT / MIDPOINTS=4.0 TO 19.0 BY 1.0 NORMAL;
RUN;

PROC UNIVARIATE DATA=LIVEREXP NORMAL PLOT;
        TITLE "DESCRIPTIVE STATISTICS FOR LIVER_WT AND DOSE";
        BY DOSE;
        VAR LIVER_WT;
        HISTOGRAM LIVER_WT / MIDPOINTS=9.0 TO 14.0 BY 0.5 NORMAL;
        HISTOGRAM LIVER_WT / MIDPOINTS=9.0 TO 14.0 BY 1.0 NORMAL;
RUN;

**TEST FOR DIFFERENCE BETWEEN VARIABLES;

PROC NPAR1WAY DATA=LIVEREXP WILCOXON;
        TITLE "NONPARAMETRIC TEST FOR REACT AND DOSE";
        CLASS DOSE;
        VAR REACT;
        EXACT WILCOXON;
RUN;

PROC TTEST DATA=LIVEREXP;
        TITLE "T-TEST FOR LIVER_WT AND DOSE";
        CLASS DOSE;
        VAR LIVER_WT;
RUN;
```