STAT 6800 Homework 2

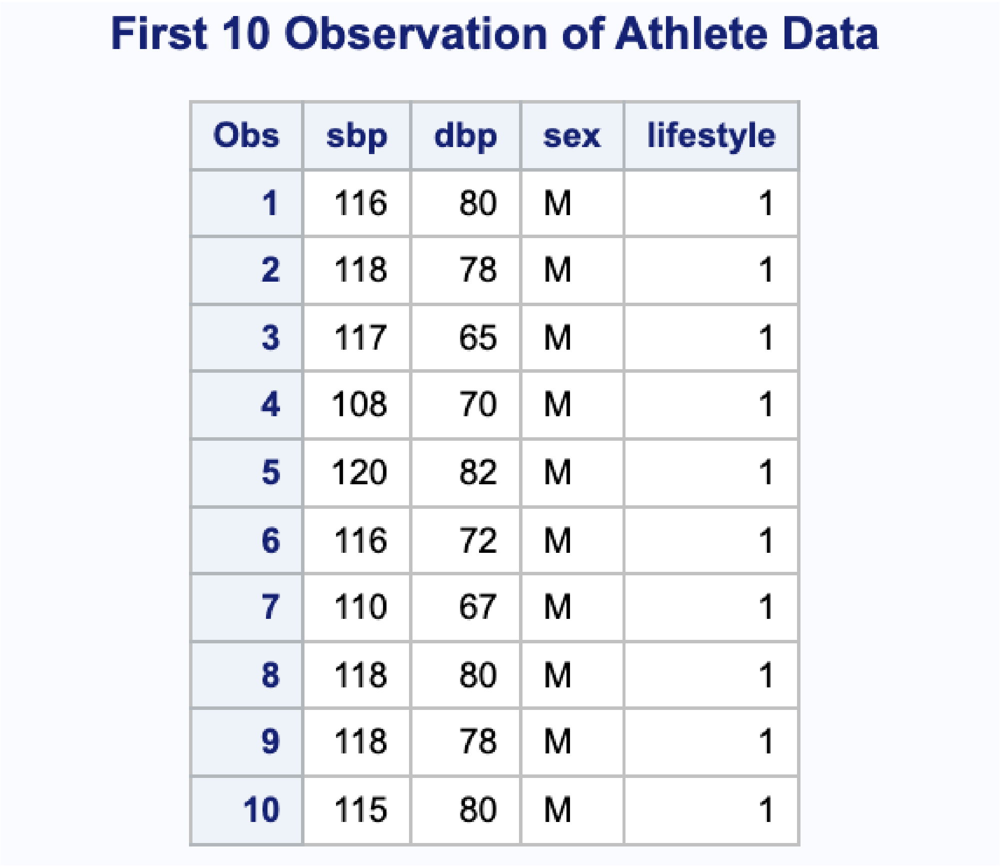
Augustine Ennin

September 2025

Q1.(a)

```
DATA athlete;
INFILE "/home/u63997979/sasuser.v94/Elliott and Morrell/athlete.dat";
INPUT sbp 1-3 dbp 6-7 sex $ 10 lifestyle 13;
RUN;

PROC PRINT DATA=athlete(OBS = 10);
TITLE "First 10 Observation of Athlete Data";
RUN;
```



First 10 Observation of Athlete Data

Obs	sbp	dbp	sex	lifestyle
1	116	80	M	1
2	118	78	M	1
3	117	65	M	1
4	108	70	M	1
5	120	82	M	1
6	116	72	M	1
7	110	67	M	1
8	118	80	M	1
9	118	78	M	1
10	115	80	M	1

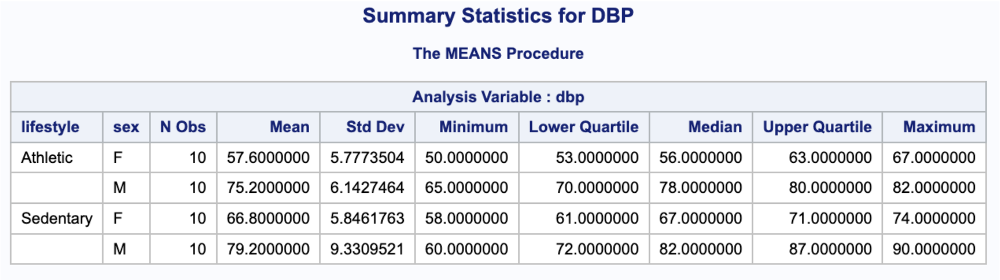
Figure 1: First 10 data output

Q1.(b)

```
PROC FORMAT;
        VALUE lifestylefmt 1 = "Athletic" 2 = "Sedentary";
```

```
RUN;

PROC MEANS DATA = athlete MEAN STD MIN Q1 MEDIAN Q3 MAX;
CLASS lifestyle sex;
VAR dbp;
FORMAT lifestyle lifestylefmt.;
TITLE "Summary Statistics for DBP";
RUN;
```



Summary Statistics for DBP

The MEANS Procedure

Analysis Variable : dbp									
lifestyle	sex	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Athletic	F	10	57.6000000	5.7773504	50.0000000	53.0000000	56.0000000	63.0000000	67.0000000
	M	10	75.2000000	6.1427464	65.0000000	70.0000000	78.0000000	80.0000000	82.0000000
Sedentary	F	10	66.8000000	5.8461763	58.0000000	61.0000000	67.0000000	71.0000000	74.0000000
	M	10	79.2000000	9.3309521	60.0000000	72.0000000	82.0000000	87.0000000	90.0000000

Figure 2: Summary statistics

Q1.(c)

```
PROC UNIVARIATE DATA = athlete NORMAL;
VAR sbp;
PROBPLOT sbp / NORMAL(MU = EST SIGMA = EST);
TITLE "Normal Probability Plot for SBP";
RUN;
```

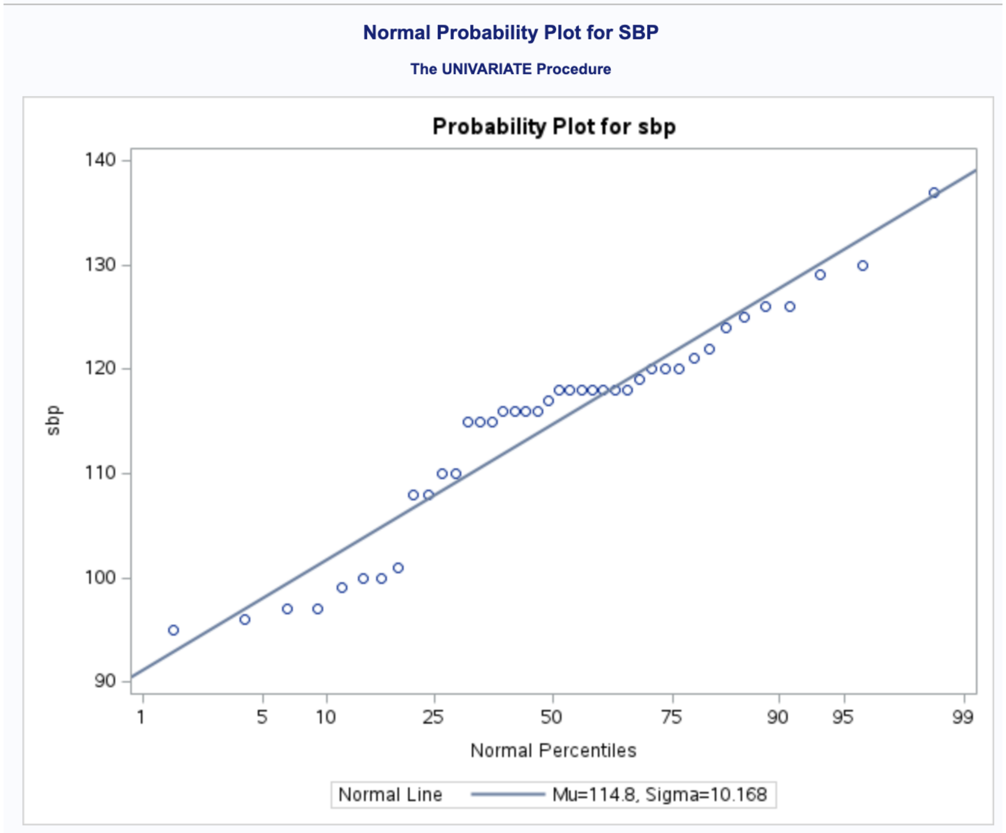
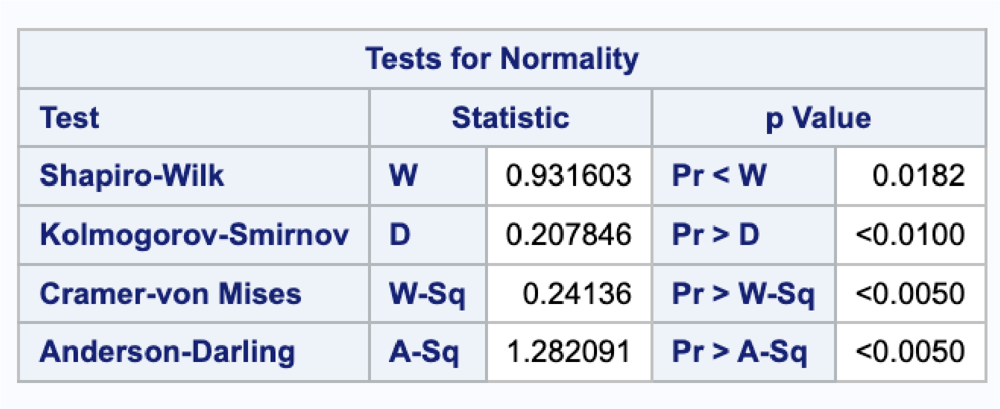


Figure 3: Probability plot



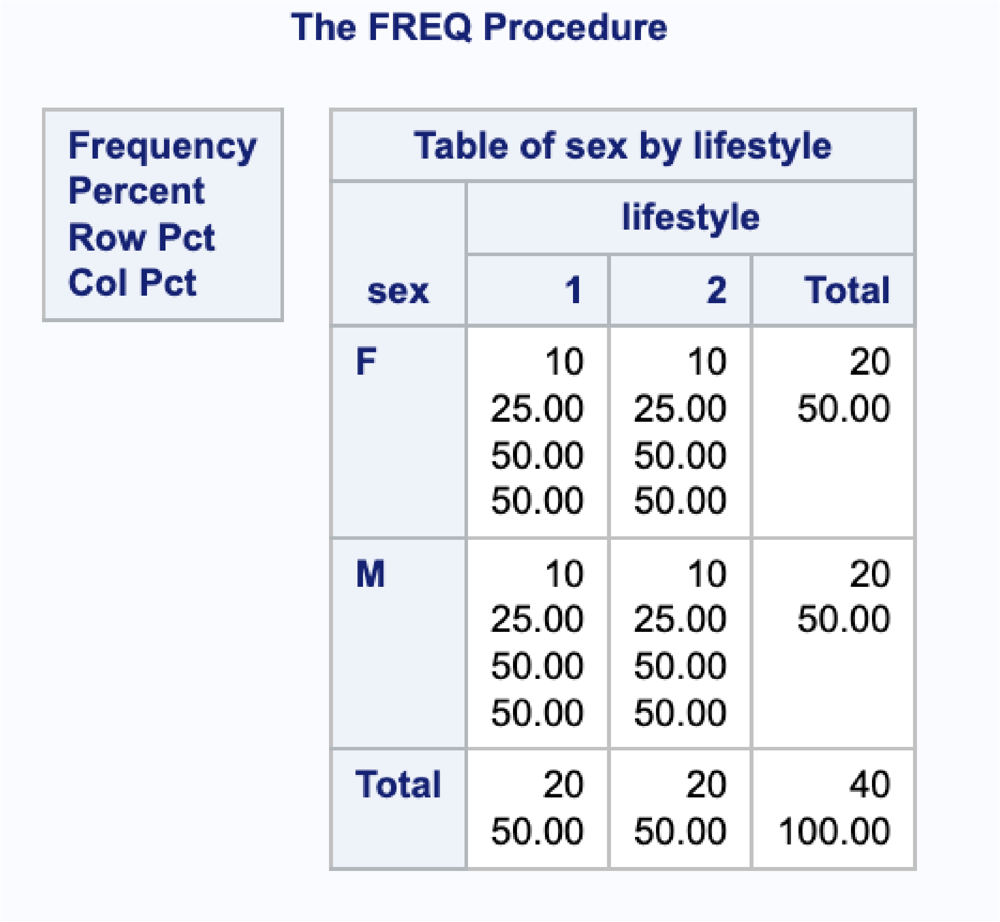
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.931603	**Pr < W**	0.0182
Kolmogorov-Smirnov	**D**	0.207846	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.24136	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	1.282091	**Pr > A-Sq**	<0.0050

Figure 4: Normality test

H_0: Systolic blood pressure is normally distributed
H_1: Systolic blood pressure is not normally distributed
From the Tests for Normality table, the Anderson–Darling test produced a $p - value < 0.005$, which is below the significance level $\alpha = 0.05$. Therefore, we reject the null hypothesis and conclude that the systolic blood pressure readings do not follow a normal distribution.

Q1.(d)

```
PROC FREQ DATA = athlete;
TABLES sex * lifestyle;
RUN;
```



The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of sex by lifestyle

	lifestyle		
sex	1	2	Total
F	10 25.00 50.00 50.00	10 25.00 50.00 50.00	20 50.00
M	10 25.00 50.00 50.00	10 25.00 50.00 50.00	20 50.00
Total	20 50.00	20 50.00	40 100.00

Figure 5: The frequency procedure

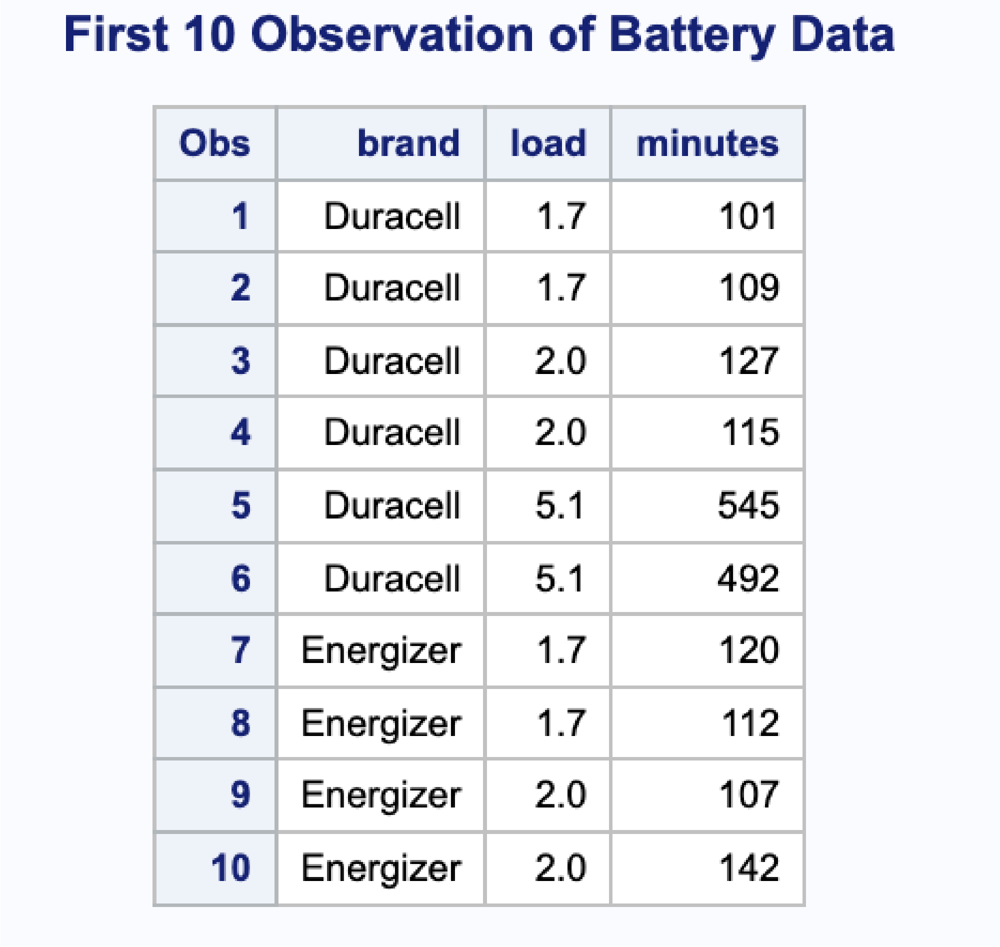
P(sex, lifestyle) = freq in cell / total observation.
Each cell has 10/40 = 0.25 joint probability. This is the same as ”Percent” value in each observation.

Q2.(a)

```
PROC FORMAT;
VALUE brandfmt 1="Duracell" 2="Energizer" 3="Rayovac" 4="Radio Shack";
RUN;

DATA battery;
INFILE "/home/u63997979/sasuser.v94/Elliott and Morrell/battery.dat";
INPUT brand 1 load 4-6 minutes 9-11;
FORMAT brand brandfmt.;
RUN;

PROC PRINT DATA=battery(OBS=10);
TITLE "First 10 Observation of Battery Data";
RUN;
```



First 10 Observation of Battery Data

Obs	brand	load	minutes
1	Duracell	1.7	101
2	Duracell	1.7	109
3	Duracell	2.0	127
4	Duracell	2.0	115
5	Duracell	5.1	545
6	Duracell	5.1	492
7	Energizer	1.7	120
8	Energizer	1.7	112
9	Energizer	2.0	107
10	Energizer	2.0	142

Figure 6: First 10 observation

Q2.(b)

```
PROC UNIVARIATE DATA = battery;
VAR minutes;
HISTOGRAM minutes;
TITLE "Histogram of Battery Life (Minutes)";
RUN;
```

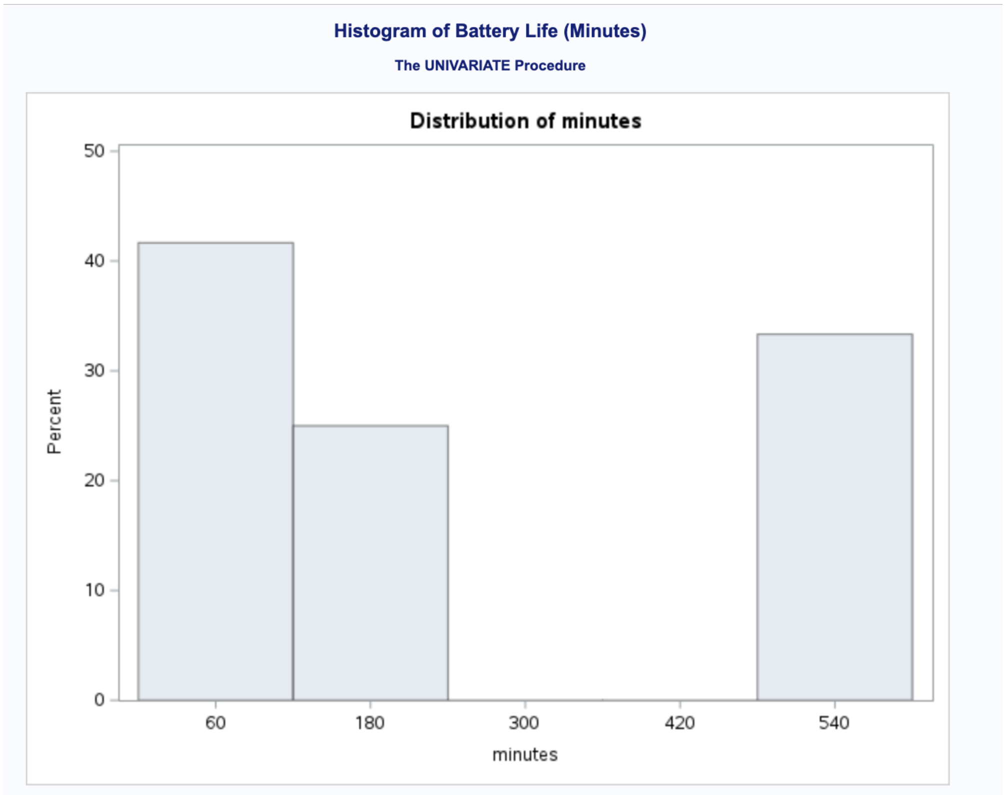
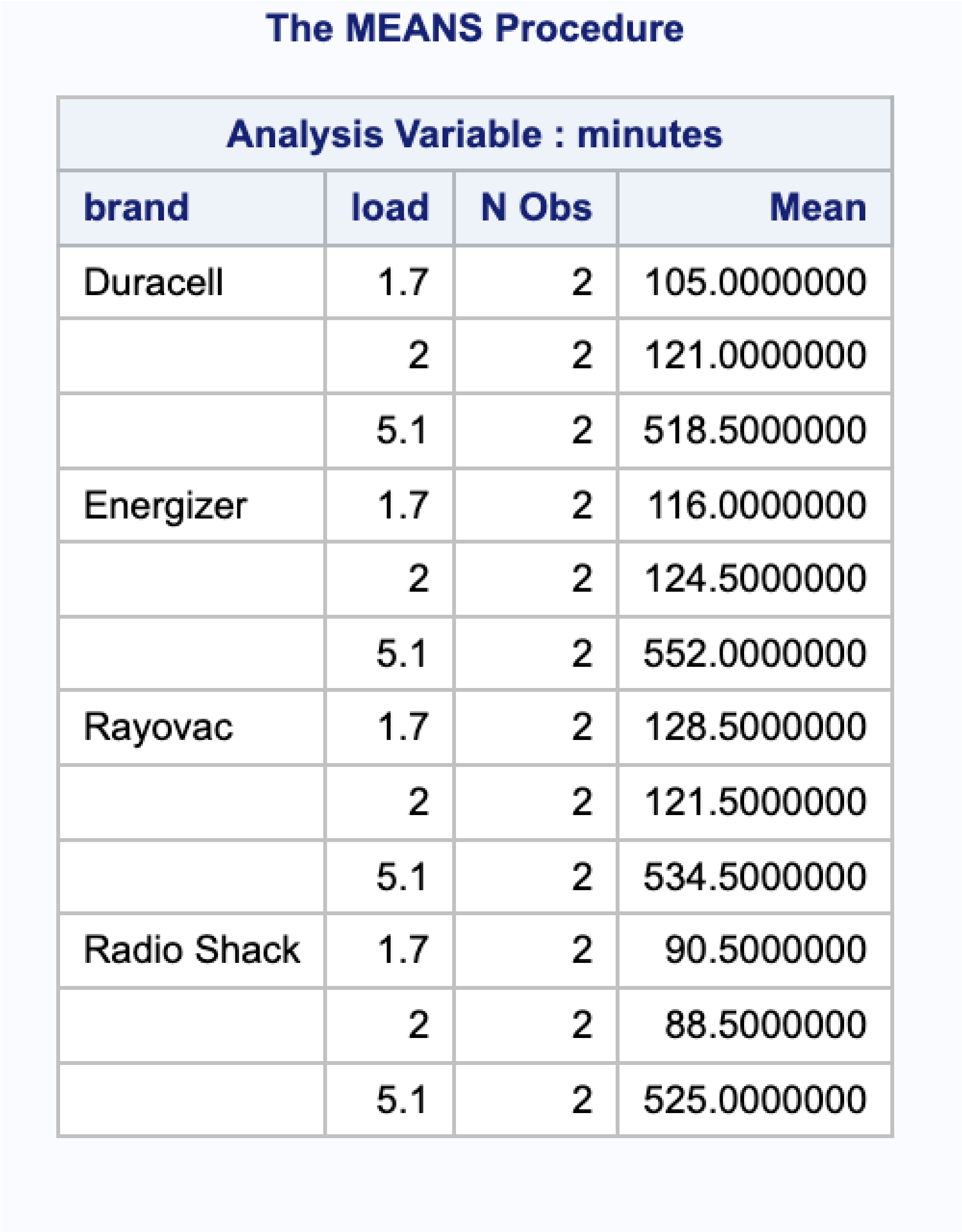


Figure 7: Histogram plot

The battery lifetime histogram has two peaks at 60 and 540 minutes with a significant gap from 180 to 540 minutes. This indicates that the minutes are bimodally distributed, suggesting two populations of batteries with different lifetimes.

Q2.(c)

```
PROC MEANS DATA = battery MEAN;
CLASS brand load;
VAR minutes;
RUN;
```



The MEANS Procedure

Analysis Variable : minutes			
brand	load	N Obs	Mean
Duracell	1.7	2	105.0000000
	2	2	121.0000000
	5.1	2	518.5000000
Energizer	1.7	2	116.0000000
	2	2	124.5000000
	5.1	2	552.0000000
Rayovac	1.7	2	128.5000000
	2	2	121.5000000
	5.1	2	534.5000000
Radio Shack	1.7	2	90.5000000
	2	2	88.5000000
	5.1	2	525.0000000

Figure 8: The frequency procedure

The PROC MEANS data supports the explanation of the histogram. The mean battery life (minutes) for load 1.7 is between 90 and 128.5 by brand, between 88 and 124.5 for load 2.0, and between 518 and 552 for load 5.1. The short peaks in the histogram at around 60 and 180 minutes are the batteries at loads 1.7 and 2.0, and the high peak at around 540 minutes is the batteries at load 5.1. Thus, the bimodal histogram is driven by the different levels of loads imposed on each type of batteries.