

Honor Statement: *You are to work individually on this Project.* All questions you have should be directed to Dr. Fridline. For this exam, you are permitted to use the textbook and any lecture notes provided. There is no collaboration of any kind with anyone! Please print and sign your name below affirming your awareness of the honor standards stated above.

Printed: _____

Signed: ENOCK KUMI ACKAAH

Problem #1:

This is a set of observations on 27 people. Several things were noted. The first variable is the hospital in which the data was collected, the second variable is the gender of the person, then a measure of the amount of iron in the blood called hemoglobin (Haemo), then packed cell volume (Pcv), white blood cell count (Wbc), the number of lymphocytes (Lympho), neutrophil (Neutro), and serum lead concentration (Lead).

A	M	13.4	39.0	4100	14	25	17
B	f	14.6	46.0	5000	15	30	20
a	F	13.5	42.0	4500	19	21	18
A	M	15	46.0	4600	23	16	18
b	m	14.6	44.0	5100	17	.	19
B	m	14	44.0	4900	20	24	19
b	f	16.4	49.0	4300	21	17	18
A	f	14.8	44.0	4400	16	26	29
a	F	15.2	46.0	4100	27	13	27
B	M	15.5	48.0	8400	34	42	36
A	M	15.2	47.0	.	26	27	22
B	F	16.9	50.0	5100	28	17	23
a	F	14.8	44.0	4700	24	20	23
A	M	16.2	45.0	5600	26	25	19
b	M	14.7	43.0	4000	23	13	17
A	F	14.7	42.0	3400	9	22	13
B	m	16.5	45.0	5400	18	32	17
B	f	15.4	45.0	6900	28	36	24
B	M	15.1	45.0	4600	.	29	17
a	M	14.2	46.0	4200	14	25	28
b	F	15.9	46.0	5200	8	34	16
B	f	16	47.0	4700	25	14	18
A	F	17.4	50.0	8600	37	.	17
A	M	14.3	43.0	5500	20	31	19
B	F	14.8	44.0	4200	15	24	19
a	f	15.5	.	4900	17	27	16
a	m	26.5	.	5200	19	32	21

The location of this dataset is: Brightspace > Course Materials > Content > Projects > Hospital.csv

- a. Load the data exactly as provided in Brightspace, do not make any changes. This will require you to figure out how to handle the missing data. Then figure out how to get SAS and SPSS to correct difficulties. F-f, A=a, and so on. Be sure to apply the appropriate formats and labels to this dataset. Print your SAS code.

a. SAS CODE

```
libname mylib "C:\Users\eka47\OneDrive - The University of Akron\Statistical  
data managemant\mylib";  
  
data mylib.hospital_data;  
  infile "C:\Users\eka47\OneDrive - The University of Akron\Statistical data  
managemant\mylib\hospital.csv" dsd firstobs=2 missover;  
  input Hospital $ Gender $ Haemo Pcv Wbc Lympho Neutro Lead;  
  
  if Gender = 'f' then Gender = 'F';  
  if Gender = 'm' then Gender = 'M';  
  if Hospital = 'a' then Hospital = 'A';  
  if Hospital = 'b' then Hospital = 'B';  
  
  label  
    Hospital = 'Hospital Code'  
    Gender = 'Gender of the Person'  
    Haemo = 'Hemoglobin Level'  
    Pcv = 'Packed Cell Volume'  
    Wbc = 'White Blood Cell Count'  
    Lympho = 'Number of Lymphocytes'  
    Neutro = 'Neutrophil Count'  
    Lead = 'Lead Concentration';  
  
run;  
  
proc print data=mylib.hospital_data label;  
run;
```

NOTE

- This code loads a dataset from a CSV file, converts certain variables to uppercase for consistency, and applies descriptive labels to each variable.
- It then prints the dataset, showing the more user-friendly labels for the variables instead of the raw variable names.

DATASET IN SAS WITH NO CHANGES AS REQUIRED

Obs	Hospital Code	Gender of the Person	Hemoglobin Level	Packed Cell Volume	White Blood Cell Count	Number of Lymphocytes	Neutrophil Count	Lead Concentration
1	A	M	13.4	39	4100	14	25	17
2	B	F	14.6	46	5000	15	30	20
3	A	F	13.5	42	4500	19	21	18
4	A	M	15.0	46	4600	23	16	18
5	B	M	14.6	44	5100	17	.	19
6	B	M	14.0	44	4900	20	24	19
7	B	F	16.4	49	4300	21	17	18
8	A	F	14.8	44	4400	16	26	29
9	A	F	15.2	46	4100	27	13	27
10	B	M	15.5	48	8400	34	42	36
11	A	M	15.2	47	.	26	27	22
12	B	F	16.9	50	5100	28	17	23
13	A	F	14.8	44	4700	24	20	23
14	A	M	16.2	45	5600	26	25	19
15	B	M	14.7	43	4000	23	13	17
16	A	F	14.7	42	3400	9	22	13
17	B	M	16.5	45	5400	18	32	17
18	B	F	15.4	45	6900	28	36	24
19	B	M	15.1	45	4600	.	29	17
20	A	M	14.2	46	4200	14	25	28
21	B	F	15.9	46	5200	8	34	16
22	B	F	16.0	47	4700	25	14	18
23	A	F	17.4	50	8600	37	.	17
24	A	M	14.3	43	5500	20	31	19

INTERPRETATION OF THE OUTPUT ABOVE









- After loading the data into SAS, the dataset contains the following variables: **Hospital Code**, **Gender of the Person**, **Hemoglobin Level**, **Packed Cell Volume**, **White Blood Cell Count**, **Number of Lymphocytes**, **Neutrophil Count**, and **Lead Concentration**.
- **Hospital Code**: The dataset appears to represent two hospitals, A and B.
- **Gender**: Both males (M) and females (F) are represented in the dataset.
- **Hemoglobin Level**: There is variability in hemoglobin levels across individuals, ranging from **13.4** to **17.4** in the visible rows.
- **Packed Cell Volume**: Most individuals have a packed cell volume between **39** and **50**, aligning with the earlier summary statistics.

- **White Blood Cell Count:** The white blood cell counts vary significantly, with some values as low as **3400** and others as high as **8800**.
- **Number of Lymphocytes:** Lymphocyte counts range from **8** to **37** in the visible data, which is quite variable.
- **Neutrophil Count:** This variable seems to have some missing values (denoted by a dot), and the counts range from **13** to **36**.
- **Lead Concentration:** Lead concentrations in the blood range from **13** to **29**, with some variation across individuals.
- Missing values exist for some variables (e.g., neutrophil count in certain observations), which will require handling for further analysis

- b. Load data in SPSS, do some simple descriptive statistics and charts/graphs for *Haemo* and *Lymphocytes*... Submit and comment on everything you find interesting. You must comment on ALL statistical output you print and submit

DATASET IN SPSS AND DESCRIPTIVE STATISTICS FOR HAEMOGLOBIN AND LYMPHOCYTES

DATAVIEW

	 hosp	 gender	 haemo	 pcv	 wbc	 lympho	 neutro	 lead	
	A	M	13.4	39	4100	14	25	17	
	B	F	14.6	46	5000	15	30	20	
	A	F	13.5	42	4500	19	21	18	
	A	M	15.0	46	4600	23	16	18	
	B	M	14.6	44	5100	17	.	19	
	B	M	14.0	44	4900	20	24	19	
	B	F	16.4	49	4300	21	17	18	
	A	F	14.8	44	4400	16	26	29	
	A	F	15.2	46	4100	27	13	27	
0	B	M	15.5	48	8400	34	42	36	
1	A	M	15.2	47	.	26	27	22	
2	B	F	16.9	50	5100	28	17	23	
3	A	F	14.8	44	4700	24	20	23	
4	A	M	16.2	45	5600	26	25	19	
5	B	M	14.7	43	4000	23	13	17	
6	A	F	14.7	42	3400	9	22	13	
7	B	M	16.5	45	5400	18	32	17	
8	B	F	15.4	45	6900	28	36	24	
9	B	M	15.1	45	4600	.	29	17	
0	A	M	14.2	46	4200	14	25	28	
1	B	F	15.9	46	5200	8	34	16	
2	B	F	16.0	47	4700	25	14	18	
3	A	F	17.4	50	8600	37	.	17	
4	A	M	14.3	43	5500	20	31	19	
5	B	F	14.8	44	4200	15	24	19	

VARIABLE VIEW

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
hosp	String	1	0	Hospital in which the data was collected	None	None	9	Left	Nominal	Input
gender	String	1	0	Gender of the person	None	None	9	Left	Nominal	Input
haemo	Numeric	4	1	Amount of iron in the blood called hemoglobin	None	None	8	Right	Scale	Input
pcv	Numeric	2	0	Packed cell volume	None	None	8	Right	Scale	Input
wbc	Numeric	4	0	White blood cell count	None	None	8	Right	Scale	Input
lympho	Numeric	2	0	The number of lymphocytes	None	None	8	Right	Nominal	Input
neutro	Numeric	2	0	Neutrophil	None	None	8	Right	Scale	Input
lead	Numeric	2	0	Serum lead concentration	None	None	8	Right	Scale	Input

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
The number of lymphocytes	26	8	37	20.88	6.930
Amount of iron in the blood called hemoglobin	27	13.4	26.5	15.596	2.3842
Valid N (listwise)	26				

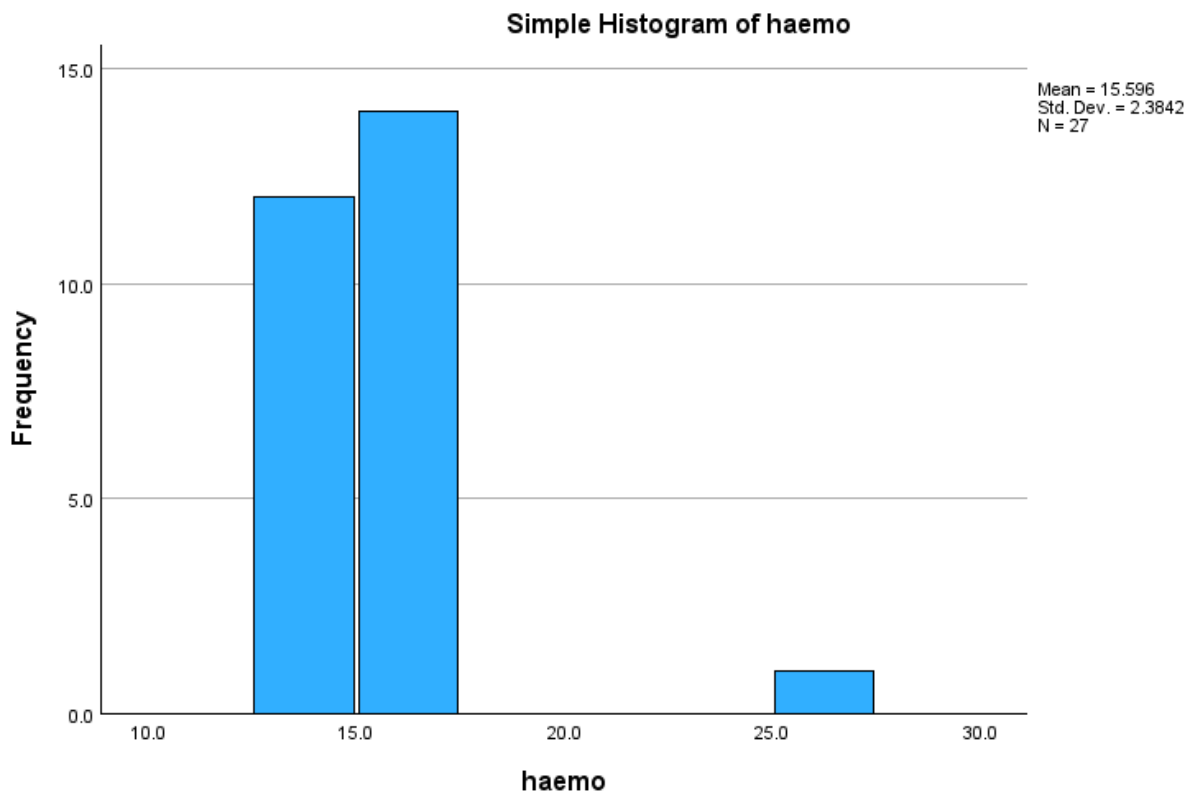
Hemoglobin:

- Sample size (N) = 27.
- The hemoglobin levels range from a minimum of **13.4** to a maximum of **26.5**.
- The mean hemoglobin level is **15.596**, indicating the average level across the sample.
- The standard deviation is **2.3842**, showing moderate variation around the mean.

Lymphocytes:

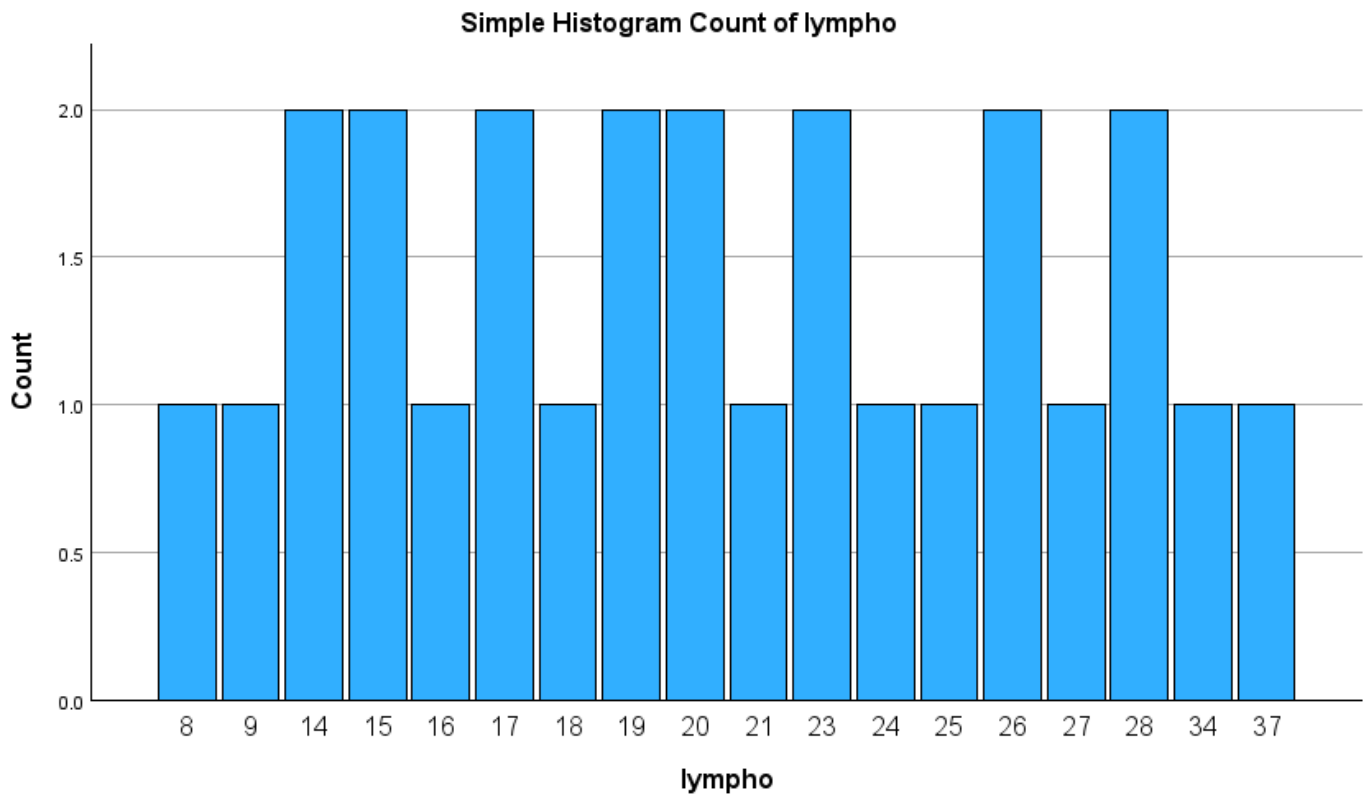
- Sample size (N) = 26, one less than the hemoglobin sample, possibly due to missing data.
- Lymphocyte counts range from **8** to **37**.
- The mean lymphocyte count is **20.88**, which is the average number of lymphocytes in the sample.
- The standard deviation is **6.930**, indicating a larger spread of data compared to hemoglobin levels, implying more variability in lymphocyte counts.

Valid N (listwise): Only **26** participants have complete data for both hemoglobin and lymphocyte counts, which explains the one missing value for the lymphocyte count.



- This histogram represents the distribution of **haemo** (hemoglobin levels) for the 27 participants in the sample.
- Many of the values are clustered between **13.4** and **17.5**, indicating that most participants have hemoglobin levels in this range. The frequency for hemoglobin levels around the mean (15.596) is notably higher, with more than half of the participants falling within this range.
- There is a small number of participants with hemoglobin levels above **25.0**, which appear as a lone bar to the right of the chart. This could suggest the presence of an outlier or a smaller subgroup with unusually high hemoglobin levels.
- The histogram shows a right-skewed (positively skewed) distribution, with most values concentrated to the left and a tail extending to the right. This skewness indicates that while most participants have lower hemoglobin levels, a few have much higher levels.
- In conclusion, this histogram suggests that hemoglobin levels are mostly concentrated around the average, with a few outliers skewing the data towards higher values.

LYMPHOCYTES



INTERPRETATION OF THE GRAPH

- This histogram suggests that the lymphocyte count is relatively spread across the observed values, with no distinct peaks or clustering, aside from the two-peak values. The distribution doesn't suggest any specific patterns, such as a normal distribution or skewness, making it relatively flat and dispersed.
- This could indicate a diverse sample population in terms of lymphocyte counts, without any strong concentration of values in particular ranges. Further analysis might explore whether this distribution changes when stratified by variables such as gender or hospital location.
- There are some outliers at higher lymphocyte counts (34 and 37) that occur once, potentially indicating less common measurements in this range.

- c. In SAS, get descriptive statistics and plots on *Haemo* and *Lympho* by using the PROC MEANS and PROC UNIVARIATE commands. Print code and supply output and plots for all the results you find interesting...please comment.

SAS CODE (PROC MEANS AND PROC UNIVARIATE)

```
Libname mylib "C:\Users\eka47\OneDrive - The University of Akron\Statistical  
data 10anagement\mylib";
```

```
data hospital_data;  
  set mylib.hospital_data;  
run;
```

```
proc means data=hospital_data mean std min max n;  
  var Haemo Lympho;  
run;
```

```
proc univariate data=hospital_data;  
  var Haemo Lympho;  
  histogram Haemo Lympho / normal;  
  inset mean std min max / position=ne;  
  qqplot Haemo Lympho;  
run;
```

HAEMOGLOBIN LEVEL

Log Results (1) Output Data (1)

The MEANS Procedure						
Variable	Label	Mean	Std Dev	Minimum	Maximum	N
Haemo	Hemoglobin Level	15.5962963	2.3841595	13.4000000	26.5000000	27
Lympho	Number of Lymphocytes	20.8846154	6.9300905	8.0000000	37.0000000	26

The UNIVARIATE Procedure			
Variable: Haemo (Hemoglobin Level)			
Moments			
N	27	Sum Weights	27
Mean	15.5962963	Sum Observations	421.1
Std Deviation	2.3841595	Variance	5.68421652
Skewness	3.89828197	Kurtosis	17.9334776
Uncorrected SS	6715.39	Corrected SS	147.78963
Coeff Variation	15.286703	Std Error Mean	0.45883171

Basic Statistical Measures			
Location		Variability	
Mean	15.59630	Std Deviation	2.38416
Median	15.10000	Variance	5.68422
Mode	14.80000	Range	13.10000
		Interquartile Range	1.40000

INTERPRETATION OF THE MOMENT AND BASIC STATISTICS OF HAEMOGLOBIN LEVEL

MEANS Procedure:

Haemo (Hemoglobin Level):

- The **mean** hemoglobin level is **15.60**.
- The **standard deviation (Std Dev)** is **2.38**, indicating some variability in the hemoglobin levels across the 27 observations.
- The **minimum** hemoglobin level is **13.40**, while the **maximum** is **26.50**, showing a range of **13.10**.

Lympho (Number of Lymphocytes):

- The **mean** lymphocyte count is **20.88**.
- The **standard deviation** is **6.93**, indicating more variability compared to hemoglobin levels.
- The **minimum** count is **8.00**, and the **maximum** is **37.00**, which gives a broader range of **29.00** for lymphocytes.
- There are 26 observations for lymphocytes, one less than the hemoglobin level data.

UNIVARIATE Procedure (Hemoglobin Level):

- This provides a deeper analysis of the hemoglobin level data.
- **N:** There are **27 observations** in total.

Mean: The mean is the same as in the MEANS procedure, **15.60**.

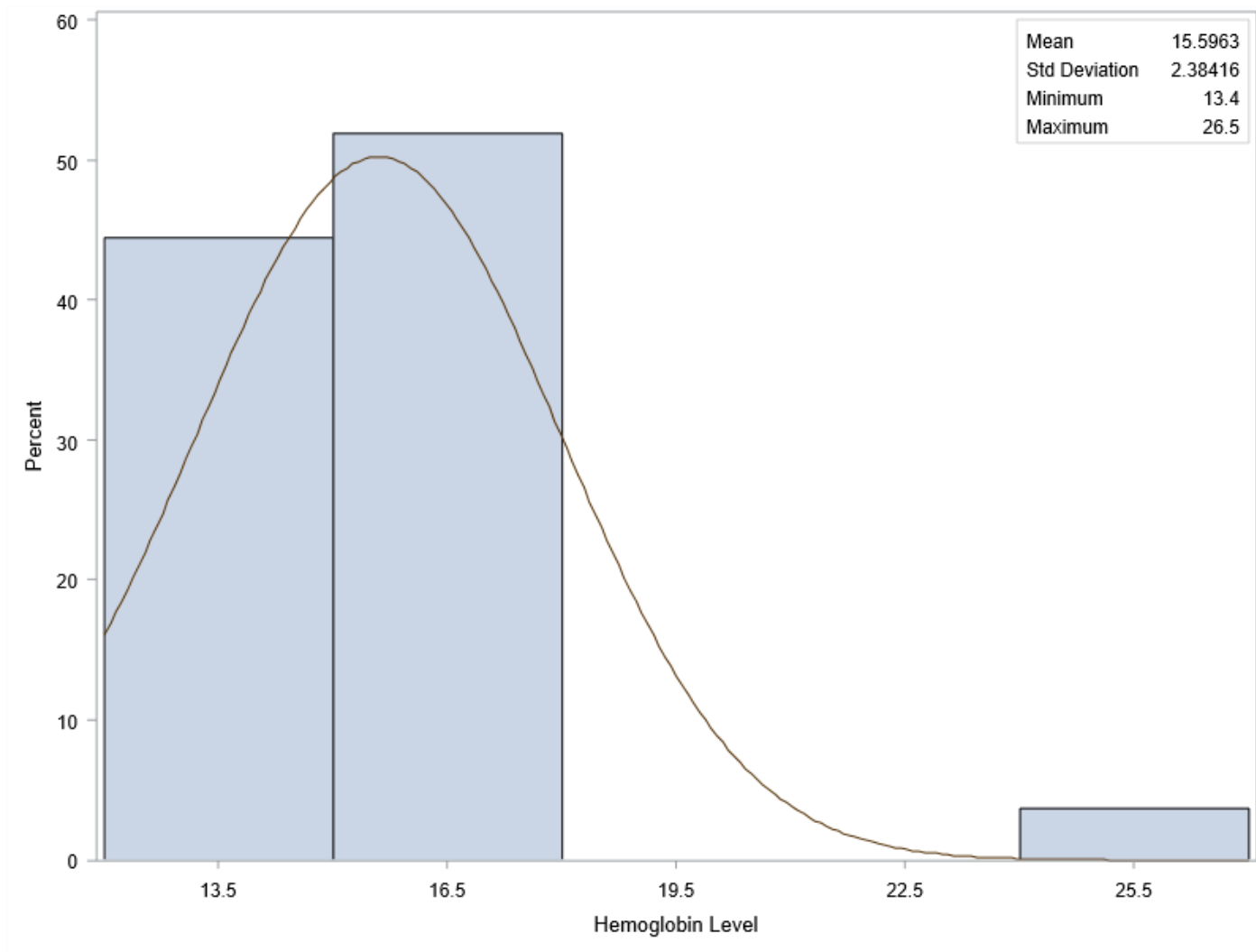
- **Standard Deviation:** **2.38**, again consistent with the MEANS output.
- **Variance:** The variance is **5.68**, which is the square of the standard deviation (confirming the calculation is correct).
- **Skewness:** The data is highly skewed to the right with a **skewness** value of **3.89**. This indicates that the hemoglobin levels are not symmetrically distributed, with a tail on the higher end.
- **Kurtosis:** The kurtosis value is extremely high at **17.93**, indicating that the distribution is leptokurtic, meaning it has more extreme values (outliers) than a normal distribution.

Basic Statistical Measures:

- **Mean:** Consistent with earlier, **15.60**.
- **Median:** The median hemoglobin level is **15.10**, which is lower than the mean, reinforcing the positive skew.
- **Mode:** The most frequent hemoglobin level is **14.80**.
- **Range:** The total range of hemoglobin levels is **13.10**.

- **Interquartile Range (IQR):** The IQR is **1.40**, indicating the spread of the middle 50% of the data is relatively small

GRAPH AND THE INTERPRETATION OF THE HAEMOGLOBIN LEVEL



- The data appears **right skewed**, with most of the hemoglobin levels concentrated between **13.4** and **16.5**. The tail extends towards higher values, indicating a few participants with much higher hemoglobin levels, up to a maximum of **26.5**.
- The histogram bars show the percentage of participants falling into each hemoglobin range. Around **50%** of the participants have hemoglobin levels between **15.5** and **16.5**, with another significant portion between **13.5** and **15**.
- In conclusion, while the distribution has some elements of normality, the right skew and outliers above **22.5** make it evident that the data is not perfectly normally distributed. Many participants

have hemoglobin levels around the mean, with a small number of participants having significantly higher values.

THE NUMBER OF LYMPHOCYTES

The UNIVARIATE Procedure Variable: Lympho (Number of Lymphocytes)			
Moments			
N	26	Sum Weights	26
Mean	20.8846154	Sum Observations	543
Std Deviation	6.93009046	Variance	48.0261538
Skewness	0.34198554	Kurtosis	0.15128538
Uncorrected SS	12541	Corrected SS	1200.65385
Coeff Variation	33.1827536	Std Error Mean	1.35910256

Basic Statistical Measures			
Location		Variability	
Mean	20.88462	Std Deviation	6.93009
Median	20.00000	Variance	48.02615
Mode	14.00000	Range	29.00000
		Interquartile Range	10.00000

COMMENTS ON THE MOMENTS AND BASIC STATISTICAL MEASURES OF THE NUMBER OF LYMPHOCYTES

- Distribution:**

The **skewness** (0.342) and **kurtosis** (0.151) are both close to zero, suggesting that the data is approximately normally distributed. There is a slight right skew, but it's minimal, and the distribution is neither particularly peaked nor flat compared to a normal distribution.

- Spread:**

The **standard deviation** of 6.93 and a **range** of 29 indicate there is quite a bit of variability in the number of lymphocytes across the sample.

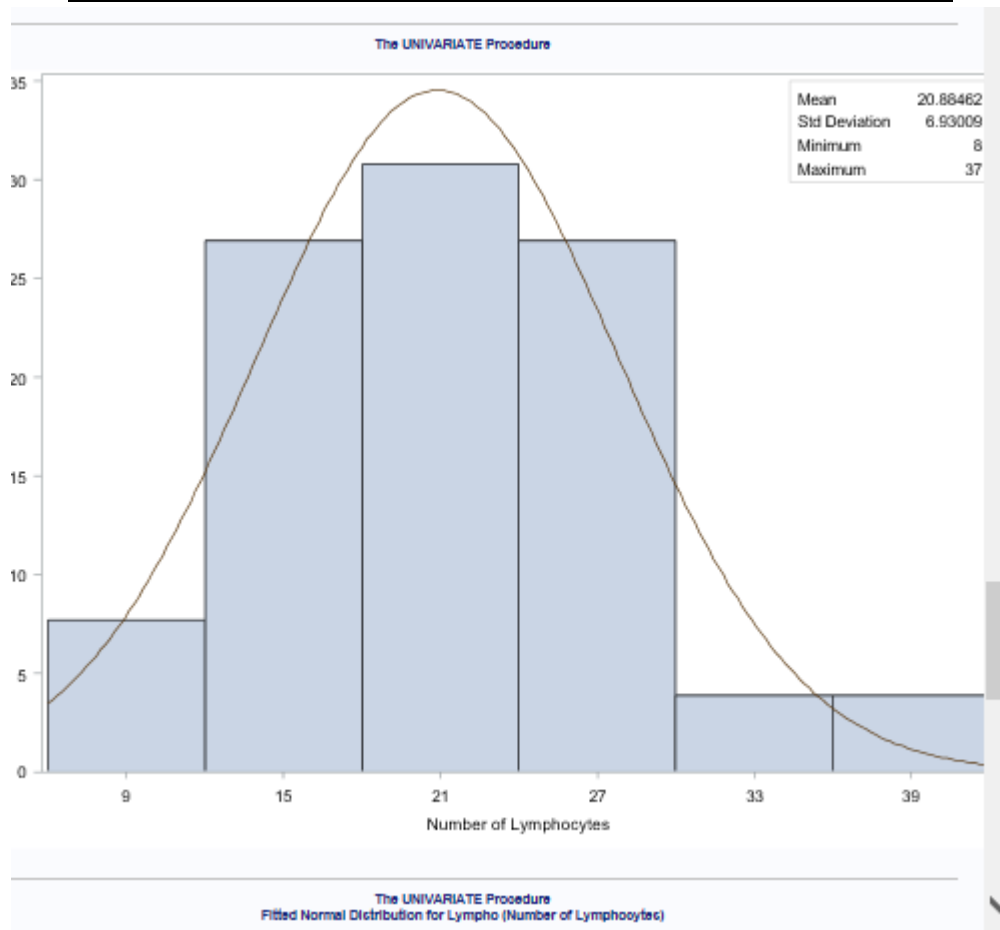
The **interquartile range** (IQR) of 10 suggests that the central 50% of the data is spread over a wide range, indicating moderate variability.

- Central Tendency:**

The **mean** (20.88) and **median** (20.00) are close, which, along with the low skewness, suggests that the distribution is symmetric, and that the data is not strongly influenced by outliers.

The **mode** (14.00) is lower than both the mean and median, indicating that although 14 is the most frequent value, many values are higher than this

GRAPH INTERPRETATION OF THE NUMBER OF LYMPHOCYTES



- The data appears to follow a **roughly normal distribution**, with a clear peak around the center. However, the histogram shows some deviations from normality, especially in the tail regions
- There is a slight **right skew**, as indicated by the few higher lymphocyte counts (above 33), pulling the distribution to the right.
- In conclusion, the data mostly follows a normal distribution but with a slight skew to the right due to the presence of a few high lymphocyte counts. Most participants have lymphocyte counts near the mean, but there are some extreme values that impact the overall distribution.

- d. *Get Pcv descriptive statistics for females and males separately by doing the following:*
 - i. *Use the subset of data command in SAS.*

SAS COMMAND IN SAS

```
libname mylib "C:\Users\eka47\OneDrive - The University of Akron\Statistical
data managemant\mylib";
```

```
data hospital_data_females;
  set mylib.hospital_data;
  if Gender = 'f';
run;
```

```
data hospital_data_males;
  set mylib.hospital_data;
  if Gender = 'm';
run;
```

```
proc means data=hospital_data_females mean std min max n;
  var Pcv;
run;
```

```
proc means data=hospital_data_males mean std min max n;
  var Pcv;
run;
```

The MEANS Procedure				
Analysis Variable : Pcv Packed Cell Volume				
Mean	Std Dev	Minimum	Maximum	N
45.7692308	2.6818478	42.0000000	50.0000000	13

The MEANS Procedure				
Analysis Variable : Pcv Packed Cell Volume				
Mean	Std Dev	Minimum	Maximum	N
44.5833333	2.3143164	39.0000000	48.0000000	12

INTERPRETATION OF THE DESCRIPTIVE STATISTICS OF PACKED CELL VOLUME

- In the first table, the **mean** PCV value is **45.77** with a **standard deviation** of **2.68** across 13 observations. The values range from **42** to **50**.
- In the second table, the **mean** is slightly lower at **44.58** with a **standard deviation** of **2.31**, across 12 observations. The range here is **39** to **48**.

- The difference in the means is small but noticeable, and the standard deviation is also lower in the second group, indicating less variability. The ranges don't overlap much, which suggests there might be some differences between these groups. The first group has a slightly higher PCV on average, while the second group has lower values, starting from 39

ii. Use the BY statement in SAS and the SPLIT command in SPSS.

BY STATEMENT IN SAS

```
libname mylib "C:\Users\eka47\OneDrive - The University of Akron\Statistical  
data managemant\mylib";
```

```
proc sort data=mylib.hospital_data;  
  by Gender;  
run;
```

```
proc means data=mylib.hospital_data mean std min max n;  
  by Gender;  
  var Pcv;  
run;
```

The MEANS Procedure				
Gender of the Person=f				
Analysis Variable : Pcv Packed Cell Volume				
Mean	Std Dev	Minimum	Maximum	N
45.7692308	2.6818478	42.0000000	50.0000000	13

Gender of the Person=m				
Analysis Variable : Pcv Packed Cell Volume				
Mean	Std Dev	Minimum	Maximum	N
44.5833333	2.3143164	39.0000000	48.0000000	12

COMPARING THE DISTRIBUTION AND VARIABILITY OF PCV BASED ON GENDER

- The average PCV is slightly higher for females (45.77) compared to males (44.58).
- The variation in PCV, as measured by standard deviation, is slightly higher for females (2.68) compared to males (2.31).
- The range of PCV values for females (42–50) is smaller than for males (39–48).

- This summary gives insight into the distribution and variability of PCV based on gender

OUTPUT OF THE SPLIT COMMAND IN SPSS

		<u>Descriptive Statistics</u>				
<u>Gender of the person</u>		<u>N</u>	<u>Minimum</u>	<u>Maximum</u>	<u>Mean</u>	<u>Std. Deviation</u>
<u>F</u>	<u>Packed cell volume</u>	<u>13</u>	<u>42</u>	<u>50</u>	<u>45.77</u>	<u>2.682</u>
	<u>Valid N (listwise)</u>	<u>13</u>				
<u>M</u>	<u>Packed cell volume</u>	<u>12</u>	<u>39</u>	<u>48</u>	<u>44.58</u>	<u>2.314</u>
	<u>Valid N (listwise)</u>	<u>12</u>				

INTERPRETATION

- **Mean Comparison:**

Females have a higher **mean packed cell volume** (45.77) compared to males (44.58). The difference is relatively small but noticeable.

- **Spread of Data:**

The **standard deviation** for females (2.682) is slightly higher than for males (2.314). This indicates that there is slightly more variability in packed cell volume among females than males.

- **Range:**

- The **range** for females is 8 units (50 - 42), and for males, it is 9 units (48 - 39). This suggests that the variability in packed cell volume is comparable across genders, though males have a slightly wider range.

- **Sample Size:**

- Both groups have relatively small sample sizes (13 for females, 12 for males), which could limit the generalizability of these results.
- Overall, females tend to have a marginally higher packed cell volume on average than males, with a slightly higher spread in the data. Both groups show similar ranges, and the differences in variability are minimal.

iii. Use the CLASS statement in SAS.

CLASS STATEMENT IN SAS

```
libname mylib "C:\Users\eka47\OneDrive - The University of Akron\Statistical  
data managemant\mylib";
```

```
proc means data=mylib.hospital_data mean std min max n;  
  class Gender;  
  var Pcv;  
run;
```

- Program 5

Analysis Variable : Pcv Packed Cell Volume						
Gender of the Person	N Obs	Mean	Std Dev	Minimum	Maximum	N
f	14	45.7692308	2.6818478	42.0000000	50.0000000	13
m	13	44.5833333	2.3143164	39.0000000	48.0000000	12

INTERPRETATION

- **Mean Difference:**

The mean **packed cell volume** for females (45.7692) is slightly higher than for males (44.5833), indicating that, on average, females in this sample have a higher PCV than males.

- **Variability:**

The **standard deviation** for females (2.6818) is slightly higher than for males (2.3143), suggesting that there is more variability in packed cell volumes among females compared to males.

- **Range:**

Females have a **range** of 8 units (50 - 42), while males have a **range** of 9 units (48 - 39). This indicates that the spread of values is similar across both groups, but slightly wider for males.

- **Sample Size:**

The number of valid observations is 14 for females and 13 for males, with one missing value in each group. This small sample size may limit the reliability of conclusions drawn from the data.