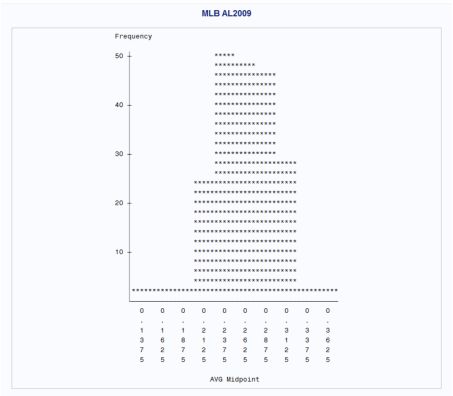
SAS Project

Jesse Annan || ID: 002708111

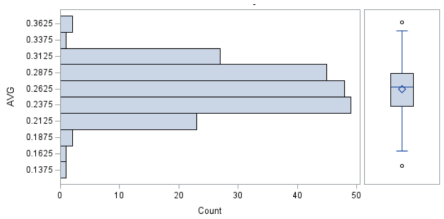
November 14, 2022

1. **Question 1: Descriptive Statistics**

   (a) MLB2009-AL.xls



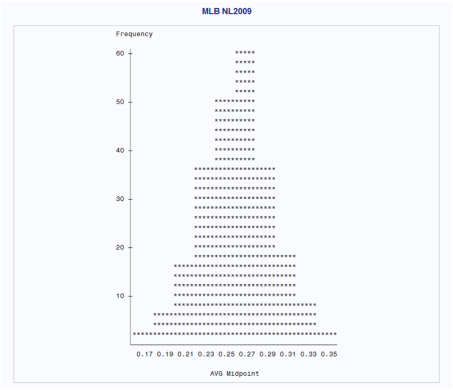
(a) Histogram



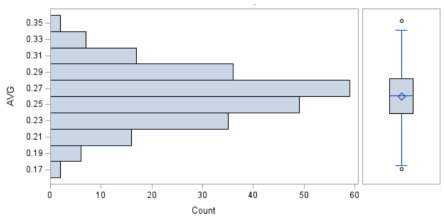
(b) Stem-and-Leaf plot

Looking at the histogram and stem-and-leaf plots we can conclude that the data points are close to the sample mean and it also looks like the data is from a normal probability distribution.

   (b) MLB2009-NL.xls



(a) Histogram



(b) Stem-and-Leaf plot

Looking at the histogram and stem-and-leaf plots we can conclude that the data is from a normal probability distribution because of the bell shape of both descriptive statistics.

2. **Question 2: Intervals and Percentage of Measurements**

(a) MLB2009-AL.xls

Basic Statistical Measures			
Location		Variability	
Mean	0.261980	Std Deviation	0.03408
Median	0.265000	Variance	0.00116
Mode	0.250000	Range	0.22100
		Interquartile Range	0.05000

(a) Histogram

	Ranges	Count	%
$\bar{x} \pm s$	(0.22790, 0.29606)	133	66.83
$\bar{x} \pm 2s$	(0.19381, 0.33015)	192	96.48
$\bar{x} \pm 3s$	(0.15973, 0.36423)	197	98.99

(b) Intervals and % of Measurements

The %s are fairly close to the empirical rule and thus we can say the data is approximately symmetric, with clustering of measurements about the midpoint of the distribution.

(b) MLB2009-NL.xls

Basic Statistical Measures			
Location		Variability	
Mean	0.260467	Std Deviation	0.03250
Median	0.261000	Variance	0.00106
Mode	0.250000	Range	0.18200
		Interquartile Range	0.04300

(a) Histogram

	Ranges	Count	%
$\bar{x} \pm s$	(0.22797, 0.29296)	160	69.87
$\bar{x} \pm 2s$	(0.19548, 0.32546)	217	94.76
$\bar{x} \pm 3s$	(0.16298, 0.35796)	229	100.00

(b) Intervals and % of Measurements

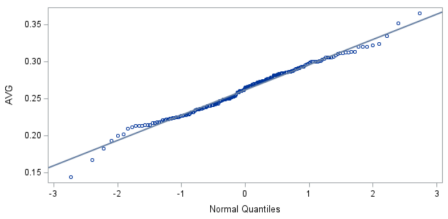
The %s are quite close to the empirical rule and thus we can say the data is approximately symmetric, with clustering of measurements about the midpoint of the distribution.

3. **Question 3: IQR/s**

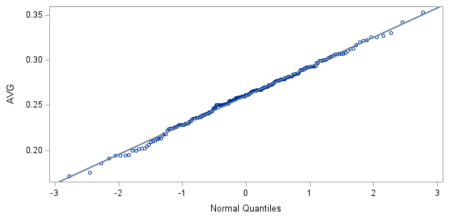
MLB2009-AL: $\frac{IQR}{s} = \frac{0.05}{0.03408} \approx 1.46695$
Since this value is a bit bigger than 1.3, we may conclude that the data are not approximately normal.

MLB2009-NL: $\frac{IQR}{s} = \frac{0.043}{0.0325} \approx 1.32324$
Since this value is approximately equal to 1.3, we have further confirmation that the data are approximately normal.

4. **Question 4: Normal Probability Plots**



(a) Normal Plot for MLB-AL2009



(b) Normal plot for MLB-NL2009

Its very clear that the AVG values fall reasonably close to the straight line for both data set and thus it suggest that both data set are approximately normally distributed.

5. **Question 5: Summary**
In this project I used SAS programming language to determine whether the data from *MLB2009-AL* and *MLB2009-NL* are from an approximately normal distribution. Checks 1 through 4 on the *MLB2009-NL* satisfied the respective rules and thus it is reasonable to believe that the data are from a normal distribution. On the other hand, *MLB2009-AL*, suggests that the data is from a normal distribution based on Checks 1, 2, and 4.
Check 3, provides doubt that the data is from a normal distribution since the ratio, $IQR/s \approx 1.46695$, is quite greater than 1.3.

1 CODE FOR MLB2009-AL.xls

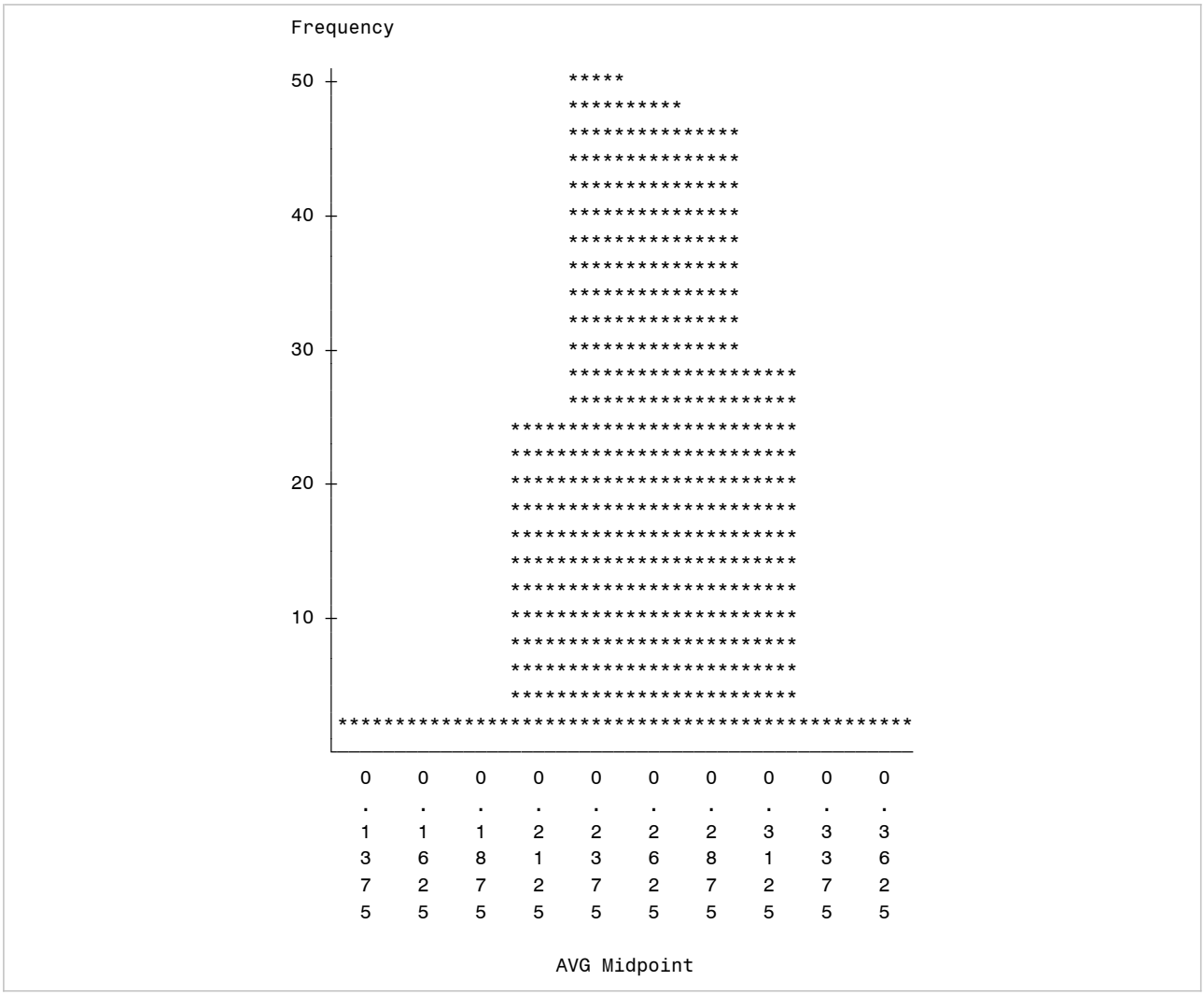
```
PROC IMPORT OUT= WORK.al2009
            DATAFILE= "C:\Users\jesse\OneDrive\Desktop\stat\MLB-
    AL2009.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

TITLE "MLB AL2009";
PROC PRINT DATA=WORK.AL2009;
run;

* drawing histogram;
PROC CHART DATA=WORK.AL2009;
  VBAR AVG /SPACE=0;
RUN;

/* mean, standard deviation, quantiles
stem-and-leaf plot, and normal plot */
PROC UNIVARIATE DATA=WORK.AL2009 NORMAL PLOT;
  VAR AVG;
RUN;
```

MLB AL2009


Frequency
50
40
30
20
10
0 0 0 0 0 0 0 0 0 0
. . . .
1 1 1 2 2 2 2 3 3 3
3 6 8 1 3 6 8 1 3 6
7 2 7 2 7 2 7 2 7 2
5 5 5 5 5 5 5 5 5 5
AVG Midpoint

MLB AL2009

The UNIVARIATE Procedure
Variable: AVG

Moments			
N	199	Sum Weights	199
Mean	0.2619799	Sum Observations	52.134
Std Deviation	0.03408426	Variance	0.00116174
Skewness	-0.0965284	Kurtosis	0.24445271
Uncorrected SS	13.888084	Corrected SS	0.23002392
Coeff Variation	13.0102587	Std Error Mean	0.00241617

Basic Statistical Measures			
Location		Variability	
Mean	0.261980	Std Deviation	0.03408
Median	0.265000	Variance	0.00116
Mode	0.250000	Range	0.22100
		Interquartile Range	0.05000

Note: The mode displayed is the smallest of 2 modes with a count of 5.

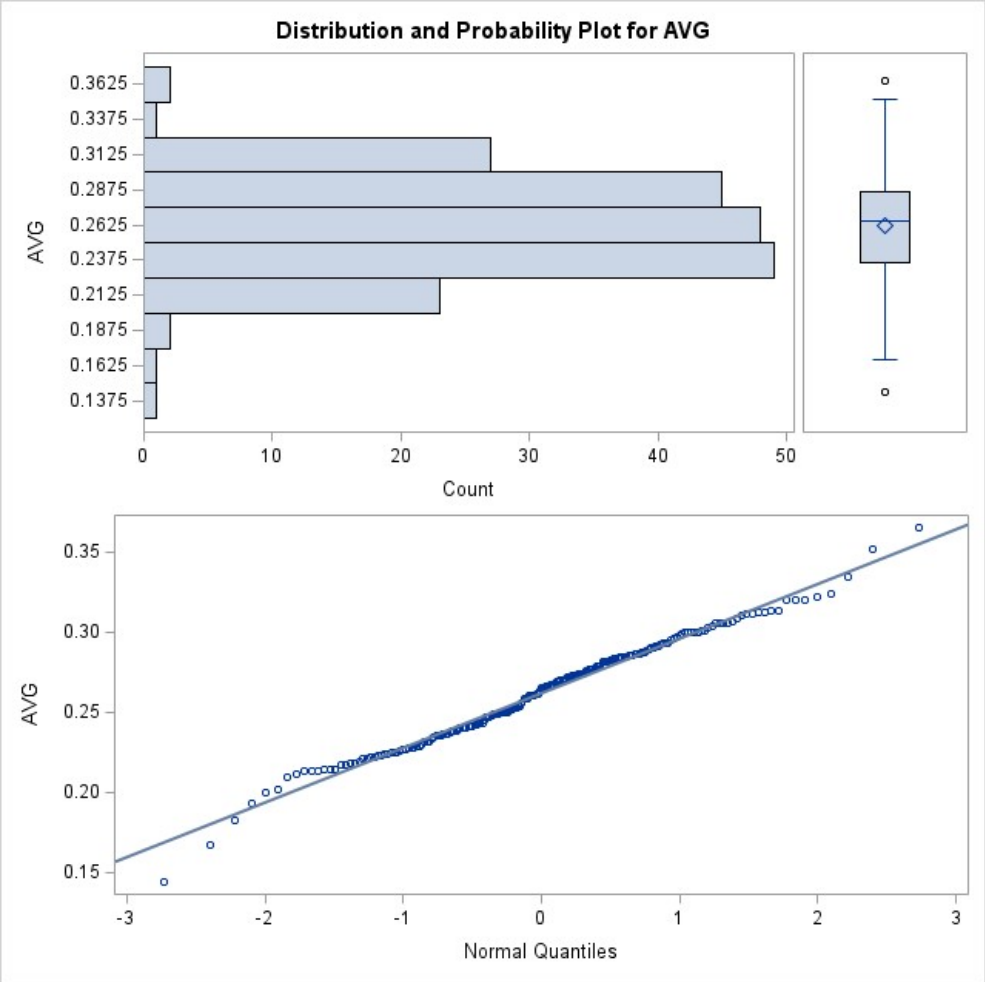
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	108.4278	Pr > |t|	<.0001
Sign	M	99.5	Pr >= |M|	<.0001
Signed Rank	S	9950	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.990281	Pr < W	0.1992
Kolmogorov-Smirnov	D	0.047869	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.093538	Pr > W-Sq	0.1393
Anderson-Darling	A-Sq	0.592867	Pr > A-Sq	0.1256

Quantiles (Definition 5)	
Level	Quantile
100% Max	0.365
99%	0.352
95%	0.313
90%	0.305
75% Q3	0.286
50% Median	0.265
25% Q1	0.236
10%	0.221
5%	0.213

1%	0.167
0% Min	0.144

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.144	195	0.322	42
0.167	163	0.324	15
0.183	188	0.334	6
0.193	134	0.352	4
0.200	189	0.365	50



2 CODE FOR MLB2009-NL.xls

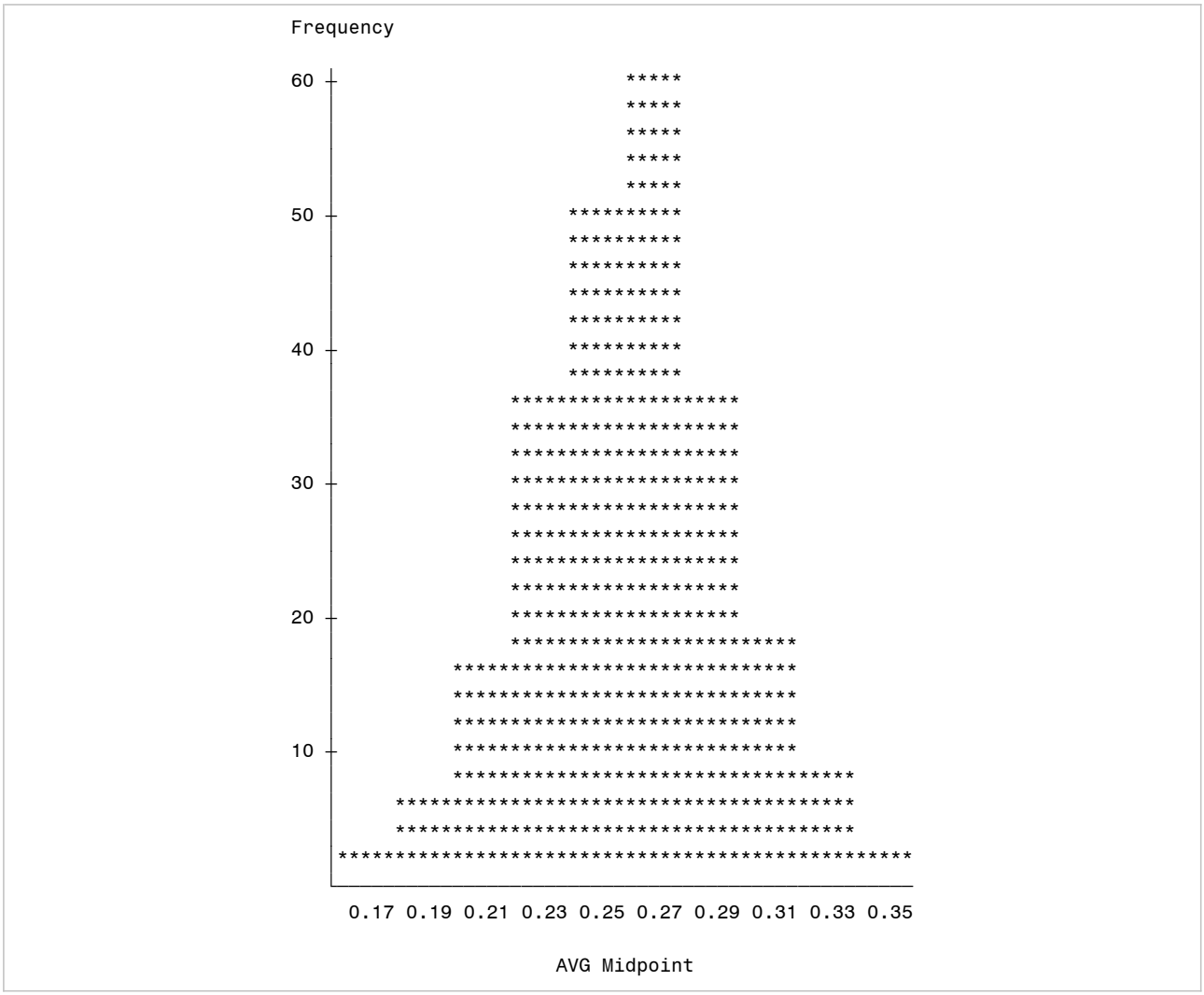
```
PROC IMPORT OUT= WORK.nl2009
            DATAFILE= "C:\Users\jesse\OneDrive\Desktop\stat\MLB-
    NL2009.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

TITLE "MLB NL2009";
PROC PRINT DATA=WORK.NL2009;
run;

* drawing histogram;
PROC CHART DATA=WORK.NL2009;
  VBAR AVG /SPACE=0;
RUN;

/* mean, standard deviation, quantiles
stem-and-leaf plot, and normal plot */
PROC UNIVARIATE DATA=WORK.NL2009 NORMAL PLOT;
  VAR AVG;
RUN;
```

MLB NL2009


Frequency
60
50
40
30
20
10
0.17 0.19 0.21 0.23 0.25 0.27 0.29 0.31 0.33 0.35
AVG Midpoint

MLB NL2009

The UNIVARIATE Procedure
Variable: AVG

Moments			
N	229	Sum Weights	229
Mean	0.26046725	Sum Observations	59.647
Std Deviation	0.03249609	Variance	0.001056
Skewness	-0.1055231	Kurtosis	0.01094625
Uncorrected SS	15.776857	Corrected SS	0.240767
Coeff Variation	12.4760739	Std Error Mean	0.0021474

Basic Statistical Measures			
Location		Variability	
Mean	0.260467	Std Deviation	0.03250
Median	0.261000	Variance	0.00106
Mode	0.250000	Range	0.18200
		Interquartile Range	0.04300

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	121.2941	Pr > |t|	<.0001
Sign	M	114.5	Pr >= |M|	<.0001
Signed Rank	S	13167.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.996656	Pr < W	0.9094
Kolmogorov-Smirnov	D	0.054908	Pr > D	0.0905
Cramer-von Mises	W-Sq	0.041086	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.252239	Pr > A-Sq	>0.2500

Quantiles (Definition 5)	
Level	Quantile
100% Max	0.353
99%	0.330
95%	0.313
90%	0.302
75% Q3	0.282
50% Median	0.261
25% Q1	0.239
10%	0.218
5%	0.202
1%	0.185
0% Min	0.171

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.171	224	0.325	120
0.175	216	0.327	26
0.185	188	0.330	23
0.191	163	0.342	22
0.194	229	0.353	158