Paper 2984-2015

SAS® for Six Sigma - An Introduction

Daniel R. Bretheim, Towers Watson, Arlington, VA

ABSTRACT

Six Sigma is a business management strategy that seeks to improve the quality of process outputs by identifying and removing the causes of defects (errors) and minimizing variability in manufacturing and business processes. Each Six Sigma project carried out within an organization follows a defined sequence of steps and has quantified financial targets. All Six Sigma project methodologies include an extensive analysis phase in which SAS® software can be applied. JMP® software is widely used for Six Sigma projects. However, this paper demonstrates how Base SAS® (and a bit of SAS/GRAPH® and SAS/STAT® software) can be used to address a wide variety of Six Sigma analysis tasks. The reader is assumed to have a basic knowledge of Six Sigma methodology. Therefore, the focus of the paper is the use of SAS code to produce outputs for analysis.

BACKGROUND

Borrowing from Kubiak and Benbow ("The Certified Six Sigma Black Belt Handbook", 2nd edition, ASQ Quality Press, 2009), the value and foundations of Six Sigma can be described as follows:

A wide range of companies have found that when the Six Sigma philosophy is fully embraced, the enterprise thrives. What is this Six Sigma philosophy? Several definitions have been proposed, with the following common threads:

- *Use of teams that are assigned well-defined projects that have direct impact on the organization's bottom line.*
- *Training in statistical thinking at all levels and providing key people with extensive training in advanced statistics and project management. These key people are designated "Black Belts."*
- *Emphasis on the DMAIC approach to problem solving: define, measure, analyze, improve, and control.*
- *A management environment that supports these initiatives as a business strategy.*

The literature is replete with examples of projects that have returned high dollar amounts to the organizations involved. Black Belts are often required to manage four projects per year for a total of $500,000-$5,000,000 in contributions to the company's bottom line.

In the first edition of their book, Kubiak and Benbow used the following to define Six Sigma:

> *Six Sigma is a fact-based, data-driven philosophy of improvement that values defect prevention over defect detection. It drives customer satisfaction and bottom-line results by reducing variation and waste, thereby promoting a competitive advantage. It applies anywhere variation and waste exist, and every employee should be involved.*

INTRODUCTION

Six Sigma projects follow one of two project methodologies that are composed of five phases each. These methodologies are referred to with the acronyms DMAIC ("duh_may_ick") and DMADV ("duh-mad_vee"). This paper will focus on DMAIC, and two of its five phases, Analyze and Control.

DMAIC:

- ***D****efine* the problem.
- ***M****easure* key aspects of the current process and collect relevant data.
- ***A****nalyze* the data to investigate and verify cause-and-effect relationships.
- ***I****mprove* or optimize the current process based upon data analysis.
- ***C****ontrol* the future state of the process to ensure that any deviations from target are corrected before they result in defects.

APPROACH

The following format will be used to describe each analytic task and how it is addressed using SAS software.

- **Issue/Objective:** A statement of the business issue or objective to be achieved by each analysis task.
- **Question:** A statement of the question to be answered.
- **SAS Procedure(s):** Identify which SAS procedure(s) are appropriate.
- **Code:** A listing of the relevant SAS code.
- **Output:** SAS output generated by the code.
- **Conclusion:** A statement that answers the question based on the output generated.

THE DATA

The examples below are based on a manufacturing scenario where the process outputs are metal plates and the quality measure is plate thickness.

The primary data set used throughout this paper consists of 2 variables and 35 observations:

Obs	Temp	Thickness
1	154	0.554
2	153	0.553
3	152	0.552
4	152	0.551
5	151	0.549
.		
.		
.		
31	147	0.542
32	147	0.542
33	146	0.541
34	146	0.540
35	145	0.538

ANALYSIS PHASE

Issue/Objective #1: When process factors that are causing excessive variation in product outputs (i.e., poor quality) are identified, measurements (i.e., data) will be collected for analysis. The data must be normally distributed in order to apply traditional statistical tests. If the data are not approximately normally distributed, alternative tests must be used.

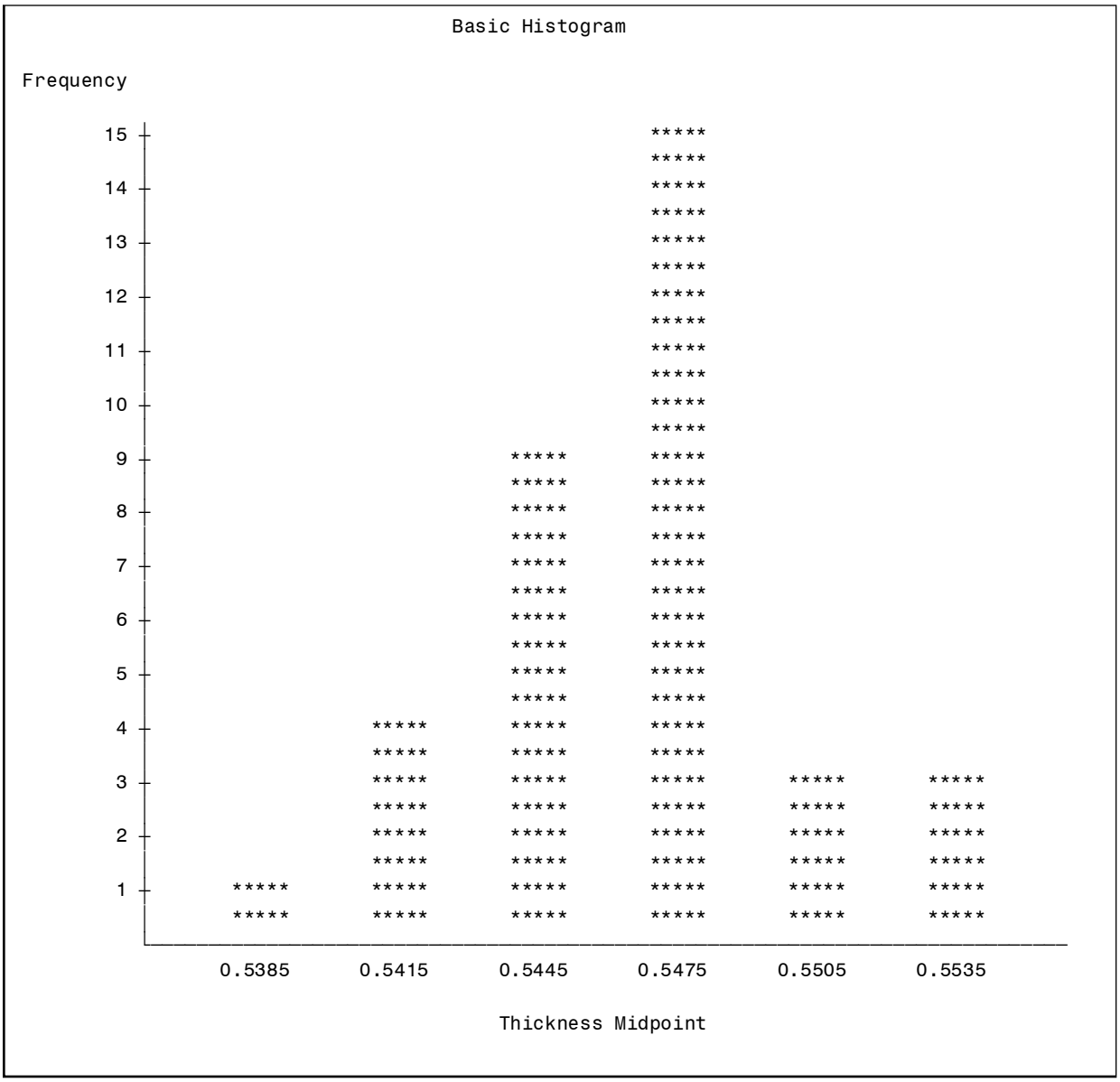
Question: Are the variables that will be analyzed approximately normally distributed or not?

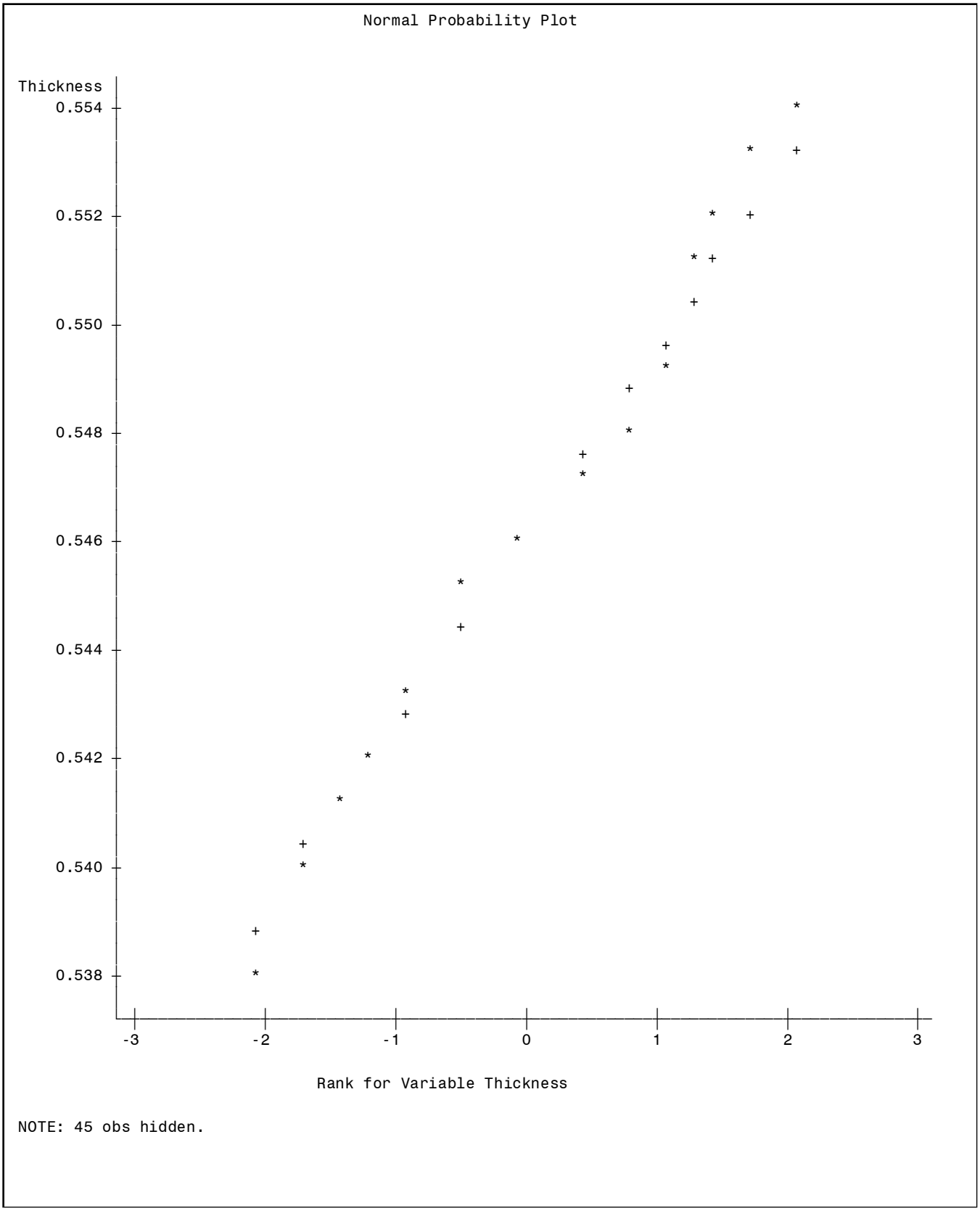
SAS Procedure(s): CHART, RANK, PLOT, MEANS, plus DATA step programming

Code:

```
** Basic Histogram ;
title 'Basic Histogram';
proc chart data=ds1;
  vbar thickness;
run;

** Normal Probability Plot ;
proc rank data=ds1 percent out=c;
  var thickness;
  ranks cage;
run;

* Cumulative probability plot of thickness ;
title 'Cumulative Probability Plot';
proc plot data=c nolegend;
  plot thickness*cage='*';
  format cage 5.1;
run;

* Calculate normal scores. ;
proc rank data=ds1 normal=blom out=r;
  var thickness;
  ranks nthickness;
run;

* Calculate mean, std, and nobs. ;
proc means data=r noprint;
  var thickness;
  output out=m mean=mean std=std;
run;

data ref;
  if _n_ = 1 then set m;
  set r;
  ethickness = mean + nthickness*std;
run;

* Produce the normal probability plot. ;
title 'Normal Probability Plot';
proc plot data=ref nolegend;
  plot thickness*nthickness='*'
ethickness*nthickness='+' /overlay;
run;
quit;
```

Output:


Basic Histogram
Frequency
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0.5385
0.5415
0.5445
0.5475
0.5505
0.5535
Thickness Midpoint



NOTE: 45 obs hidden.

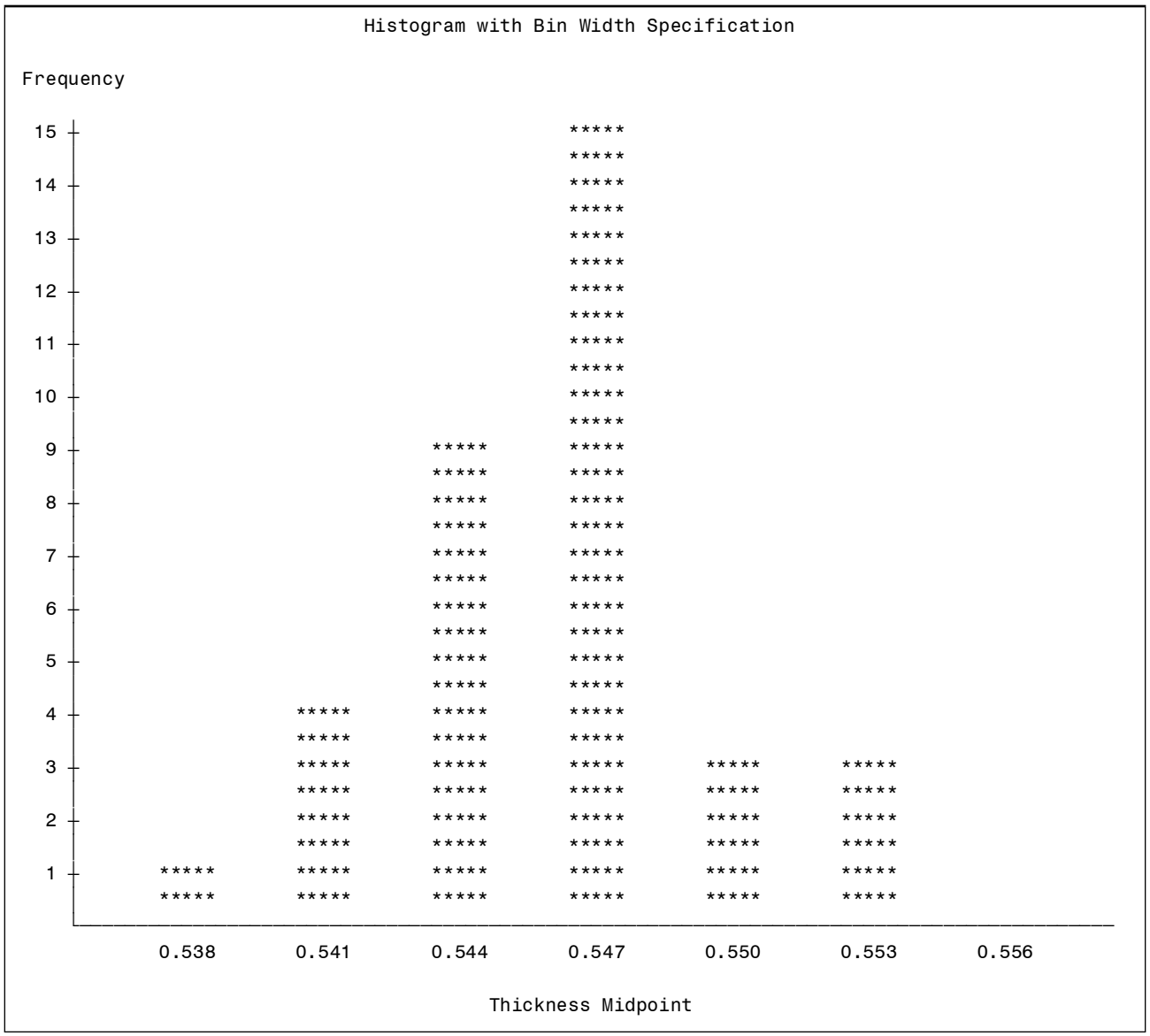
Conclusion: Yes, based on the shapes of the histogram and normal probability plot above, we can conclude that the Thickness variable is approximately normally distributed.

Optional Approach

Code:

```
** Enhanced version - Histogram with Bin Width Specification ;
proc chart data=ds1;
  vbar thickness / midpoints=.538 to .556 by .003;
run;
```

Enhanced Output:



Issue/Objective #2: Highly correlated variables may indicate a relationship of factors within process that have implications for the quality of the process outputs. Identifying such relationships is an important initial step in the Analyze phase.

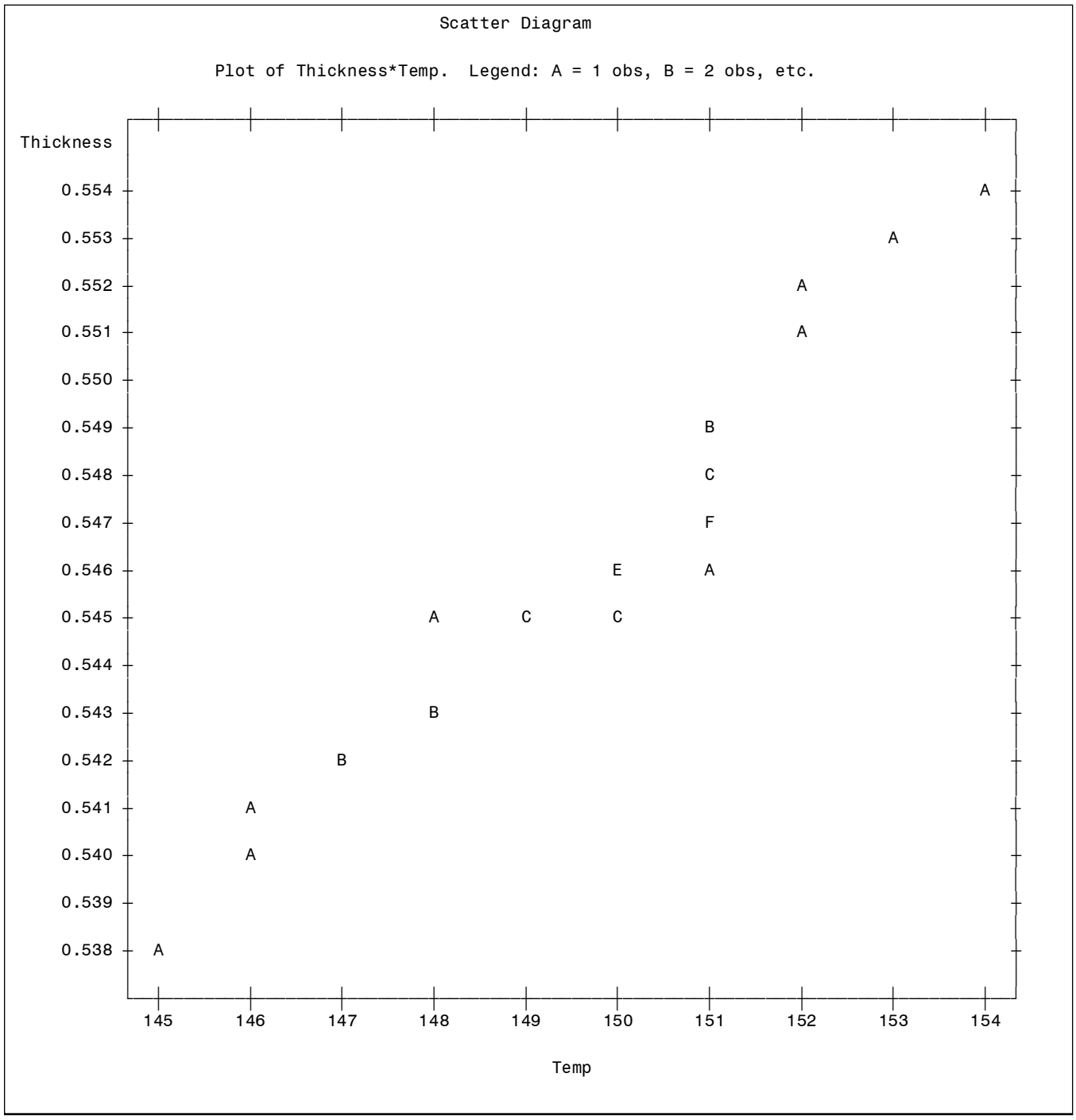
Question: What is the correlation between Temperature and Thickness?

SAS Procedure(s): PLOT, CORR

Code:

```
title 'Scatter Diagram';
proc plot data=ds1;
  plot thickness*temp / box;
run;
quit;

title 'Correlation Coefficient';
ods html;
ods graphics on;
proc corr data=ds1 plots=matrix;
  var thickness temp;
run;
ods graphics off;
ods html close;
```

Output:


Scatter Diagram
Plot of Thickness*Temp. Legend: A = 1 obs, B = 2 obs, etc.
Thickness
0.554
0.553
0.552
0.551
0.550
0.549
0.548
0.547
0.546
0.545
0.544
0.543
0.542
0.541
0.540
0.539
0.538
145
146
147
148
149
150
151
152
153
154
Temp

Correlation Coefficient

The CORR Procedure

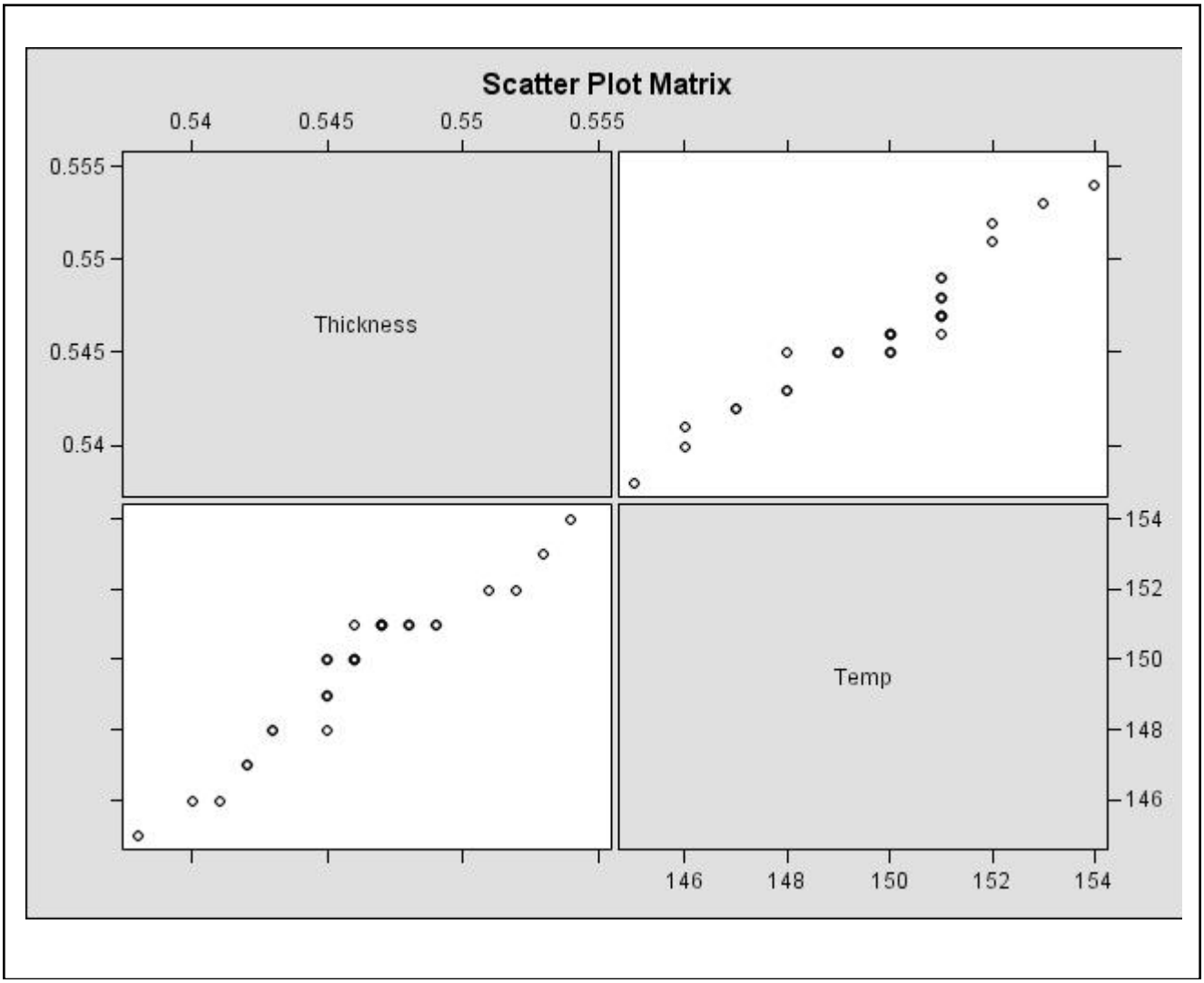
2 Variables:	Thickness Temp

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Thickness	35	0.54611	0.00339	19.11400	0.53800	0.55400
Temp	35	149.85714	1.98735	5245	145.00000	154.00000

Pearson Correlation Coefficients, N = 35
Prob > |r| under H0: Rho=0

	Thickness	Temp
Thickness	1.00000	0.95321 <.0001
Temp	0.95321 <.0001	1.00000



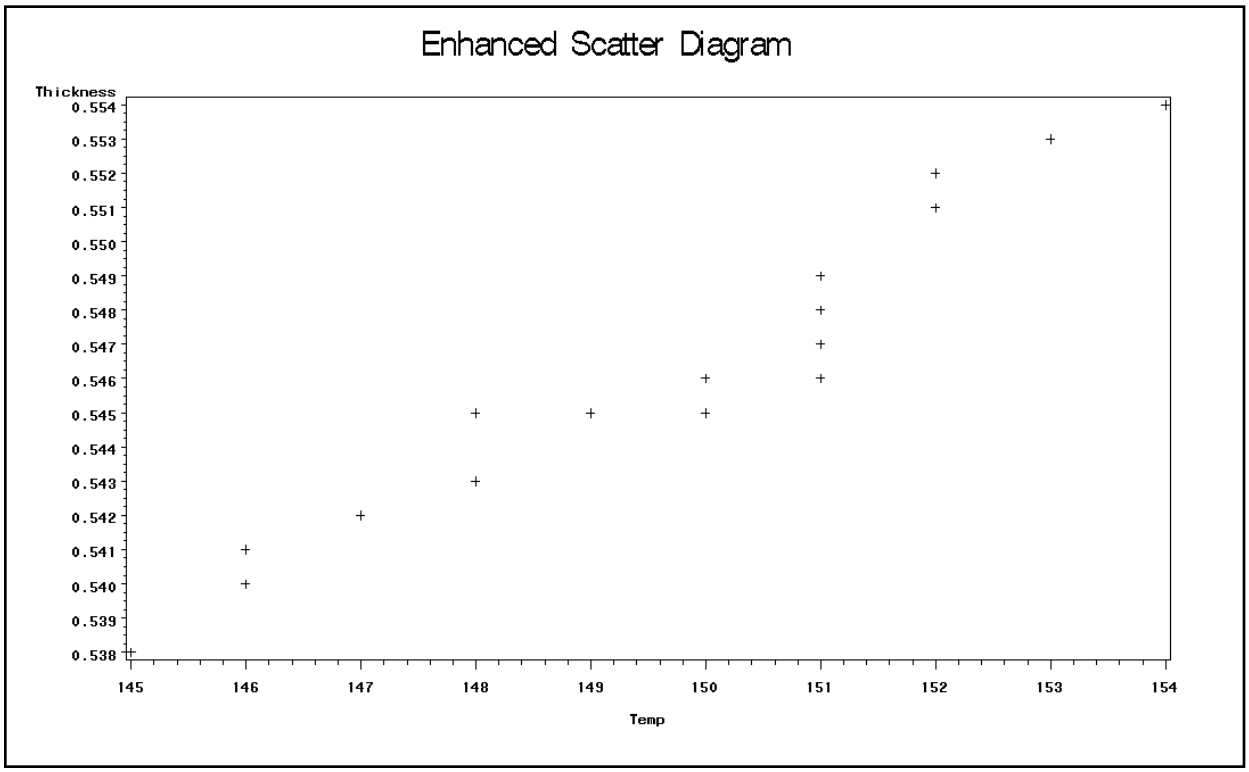
Conclusion: A correlation coefficient of .95321 indicates that Temperature and Thickness are highly positively correlated.

Optional Approach

Code:

```
title 'Enhanced Scatter Diagram';
symbol;
proc gplot data=ds1;
  plot thickness*temp;
run;
quit;
```

Enhanced Output:



Issue/Objective #3: Making a quick comparison across different levels of a process input (in this case the temperature) can provide initial insight into potential sources of variation in the process outputs. The BOXPLOT procedure available in SAS/STAT creates side-by-side box-and-whisker plots of measurements organized by groups (temperature). A box-and-whisker plot displays the mean, quartiles, and minimum and maximum observations for a group.

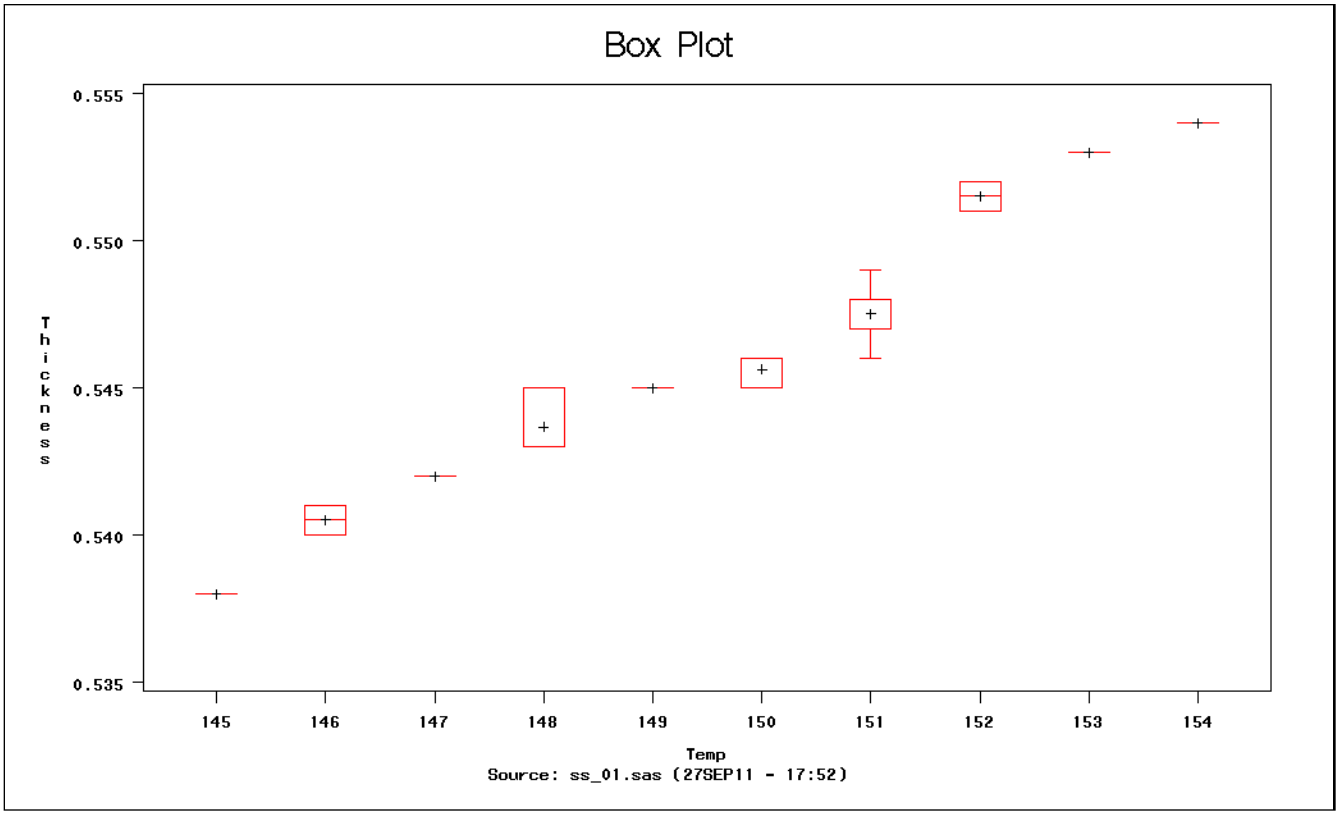
Question: How variable is Thickness at each Temperature level?

SAS Procedure(s): BOXPLOT [note the required SORT of the group variable (temp) that precedes the procedure]

Code:

```
proc sort data=ds1;
  by temp;
run;

title 'Box Plot';
proc boxplot data=ds1;
  plot thickness*temp;
run;
```

Output:



Conclusion: For Temperature levels with more than a single value for Thickness, there is minimal variation in plate Thickness.

Issue/Objective #4: Generate the maximum amount of information about the analysis variables with the least amount of code.

Question: What is the easiest way in SAS to generate a lot of information about the analysis variables?

SAS Procedure(s): UNIVARIATE (with the normal and plots options)

Code:

```
proc univariate data=ds1 normal plots;
  var thickness;
run;
```

Output:

The UNIVARIATE Procedure
Variable: Thickness

Moments

N	35	Sum Weights	35
Mean	0.54611429	Sum Observations	19.114
Std Deviation	0.00339352	Variance	0.00001152
Skewness	0.08277938	Kurtosis	0.75243633
Uncorrected SS	10.43882	Corrected SS	0.00039154
Coeff Variation	0.62139343	Std Error Mean	0.00057361

Basic Statistical Measures

Location		Variability	
Mean	0.546114	Std Deviation	0.00339
Median	0.546000	Variance	0.0000115
Mode	0.545000	Range	0.01600
		Interquartile Range	0.00300

Tests for Location: Mu0=0

Test		-Statistic-	-----p Value------
Student's t	t	952.0667	Pr > |t| <.0001
Sign	M	17.5	Pr >= |M| <.0001
Signed Rank	S	315	Pr >= |S| <.0001

Tests for Normality

Test		--Statistic---	-----p Value------
Shapiro-Wilk	W	0.961979	Pr < W 0.2620
Kolmogorov-Smirnov	D	0.171321	Pr > D <0.0100
Cramer-von Mises	W-Sq	0.134983	Pr > W-Sq 0.0376
Anderson-Darling	A-Sq	0.690889	Pr > A-Sq 0.0686

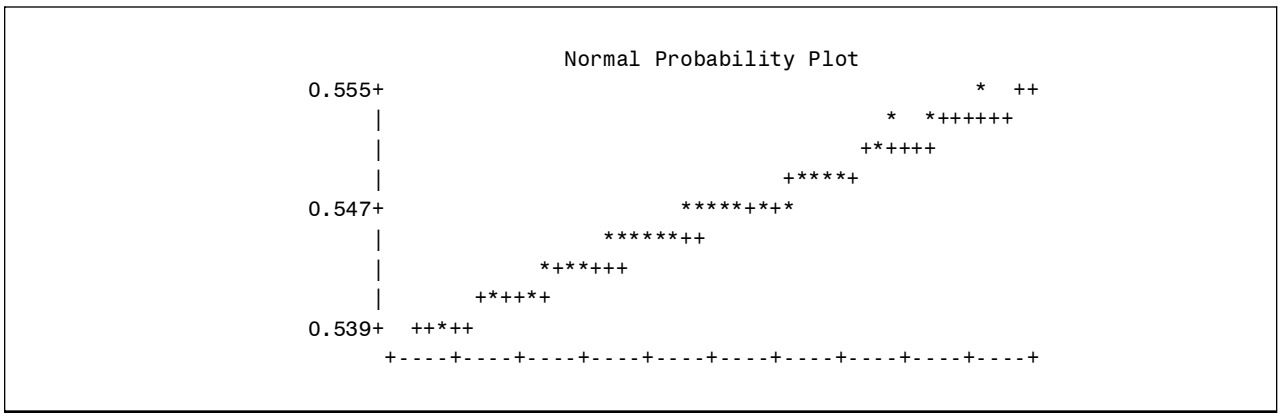
Quantiles (Definition 5)

Quantile	Estimate
100% Max	0.554
99%	0.554
95%	0.553
90%	0.551
75% Q3	0.548
50% Median	0.546
25% Q1	0.545
10%	0.542
5%	0.540
1%	0.538
0% Min	0.538

Extreme Observations

-----Lowest----		----Highest----	
Value	Obs	Value	Obs
0.538	1	0.549	21
0.540	3	0.551	33
0.541	2	0.552	32
0.542	5	0.553	34
0.542	4	0.554	35

```
Stem Leaf                     #             Boxplot
 554 0                        1                0
 552 00                       2                0
 550 0                        1                |
 548 00000                    5               +-----+
 546 000000000000            12               *--+--*
 544 0000000                  7               +-----+
 542 0000                     4                  |
 540 00                       2                0
 538 0                        1                0
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-3
```



Conclusion: A PROC UNIVARIATE with two statements/options produces most of the basic information that was covered in the first three sections above: descriptive statistics, tests for normality, outliers, stem-and-leaf plot, box plot, and normal probability plot. That's a great example of "one-stop-shopping".

Issue/Objective #5: Process owners usually want to know the limits of expected variation in process outputs.

Question: What is the expected variation, based on the mean +/- 3 standard deviations?

SAS Procedure(s): MEANS, plus DATA step programming

Code:

```
proc means data=ds1;
  var temp thickness;
  output out=stats mean= x_bar_temp x_bar_thick stddev= sd_temp sd_thick ;
run;

data expected;
  set stats;
  hi_temp = x_bar_temp + (3*sd_temp);
  lo_temp = x_bar_temp - (3*sd_temp);
  hi_thick = x_bar_thick + (3*sd_thick);
  lo_thick = x_bar_thick - (3*sd_thick);
run;

* Print for display ;
title 'Expected Variation';
proc print data=expected;
run;
```

Output:

Expected Variation

Obs	_FREQ_	x_bar_ temp	x_bar_ thick	sd_temp	sd_thick	hi_temp	lo_temp	hi_thick	lo_thick
1	35	149.857	0.54611	1.98735	.003393518	155.819	143.895	0.55629	0.53593

Conclusion: The highest point of expected variation for Temperature is 155.8 and the lowest point is 143.9. The highest point of expected variation for Thickness is .556 and the lowest point is .536.

Issue/Objective #6: If there are two methods of tapering product thickness, where the methods are designed to produce the same results, you can compare mean thickness associated with the two methods to see if they are statistically equivalent using a one sample T-Test. We know that the old method's mean is 0.012576

Question: Is the old method's average thickness the same as new method's average thickness? [This is the null hypothesis.]

SAS Procedure(s): TTEST

Data: New method thickness measurements

Obs	Thickness
1	0.009
2	0.010
3	0.011
4	0.011
5	0.010
6	0.011
7	0.011
8	0.013
9	0.008
10	0.012
11	0.010
12	0.013
13	0.014
14	0.012
15	0.009
16	0.014
17	0.011
18	0.015
19	0.011
20	0.012
21	0.015
22	0.011
23	0.011
24	0.012
25	0.008

Code:

```
title 'h0 = population mean';
/* The old method average thickness = 0.012576 */
proc ttest data=ds5 alpha=.05 h0=0.012576;
run;
```

Output:

h0 = population mean

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
Thickness	25	0.0106	0.0114	0.0121	0.0015	0.0019	0.0027	0.0004

T-Tests

Variable	DF	t Value	Pr > |t|
Thickness	24	-3.18	0.0040

Conclusion: Based on a test statistic of -3.18 (compared to a critical value of +/- 2.064), we reject the null hypothesis. We can conclude with 95 percent confidence that the old method average is not equal to the new method average, i.e., the old method and new method are producing product of different thickness.

Issue/Objective #7: In an ongoing effort to reduce variation in the process, a question has arisen about whether the manufacturing site may be influencing the defect rate. There are three sites and two types of defects that are measured. The Chi Square test for independence can be used to assess whether one or more of the three test sites may be influencing the defect rate of Type A or Type B.

Question: Are test site and defect type independent of each other?

SAS Procedure(s): FREQ

Data:

Obs	Type	Site
1	A	1
2	A	1
3	A	1
4	A	1
5	A	1
6	A	1
7	A	1
8	A	1
9	A	2
10	A	2
.		
.		
.		
42	B	2
43	B	3
44	B	3
45	B	3
46	B	3
47	B	3
48	B	3
49	B	3
50	B	3

Code:

```
title "Note the use of the 'expected' and 'chisq' options";
proc freq data=ds6;
  tables Type*Site / expected chisq;
run;
```

Output:

The FREQ Procedure

Table of Type by Site

Type	Site			
Frequency / Expected / Percent / Row Pct / Col Pct	1	2	3	Total
A	8	8	7	23
	7.82	8.28	6.9	
	16.00	16.00	14.00	46.00
	34.78	34.78	30.43	
	47.06	44.44	46.67	
B	9	10	8	27
	9.18	9.72	8.1	
	18.00	20.00	16.00	54.00
	33.33	37.04	29.63	
	52.94	55.56	53.33	
Total	17	18	15	50
	34.00	36.00	30.00	100.00

Statistics for Table of Type by Site

Statistic	DF	Value	Prob
Chi-Square	2	0.0279	0.9862
Likelihood Ratio Chi-Square	2	0.0279	0.9861
Mantel-Haenszel Chi-Square	1	0.0008	0.9776
Phi Coefficient		0.0236	
Contingency Coefficient		0.0236	
Cramer's V		0.0236	

Sample Size = 50

Conclusion: The Chi Square test for independence compares the observed values to the expected values. The observed versus expected values from the grid above are:

- Type A: 8, 7.82 8, 8.28 7, 6.9
- Type B: 9, 9.18 10, 9.72 8, 8.1

The Chi Square statistics of 0.0279 versus a critical value of 5.99 means that we fail to reject the null hypothesis that the Type A reject rate is equal to the Type B reject rate regardless of test site.

Issue/Objective #8: There is some question as to whether the three machines involved in the process are affecting plate thickness in varying degrees. We can compare the machines' performance using a one-way ANOVA.

Question: Are any of the machines significantly affecting the average material thickness?

SAS Procedure(s): ANOVA

Data:

Obs	Machine	Thickness
1	1	0.546
2	1	0.526
3	1	0.587
4	1	0.563
5	2	0.573
6	2	0.592
7	2	0.571
8	2	0.556
9	3	0.573
10	3	0.570
11	3	0.527
12	3	0.572

Code:

```
proc anova data=ds7;
  class Machine;
  model Thickness=Machine;
run;
quit;
```

Output:

The ANOVA Procedure

Dependent Variable: Thickness

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.00065000	0.00032500	0.70	0.5206
Error	9	0.00416400	0.00046267		
Corrected Total	11	0.00481400			

R-Square	Coeff Var	Root MSE	Thickness Mean
0.135023	3.820548	0.021510	0.563000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Machine	2	0.00065000	0.00032500	0.70	0.5206

Conclusion: Based on an F critical value of 4.26, versus the F calculated value of 0.70, we fail to reject the null hypothesis that the average material thickness produced by machines 1, 2, and 3 are equal. This finding tells us that none of the machines are contributing to significant variation in plate thickness.

CONTROL PHASE

After identifying the sources of unacceptable variation and modifying the process to reduce variation to an acceptable level, we move into the control phase where the process is monitored to ensure that output quality remains within acceptable limits. The XmR ("individual X moving R") charts are a tool for the on-going monitoring of key process parameters.

Issue/Objective #9: Individual outputs can be plotted on the X chart to see whether the process parameter of interest (e.g., thickness) falls within acceptable levels of variation.

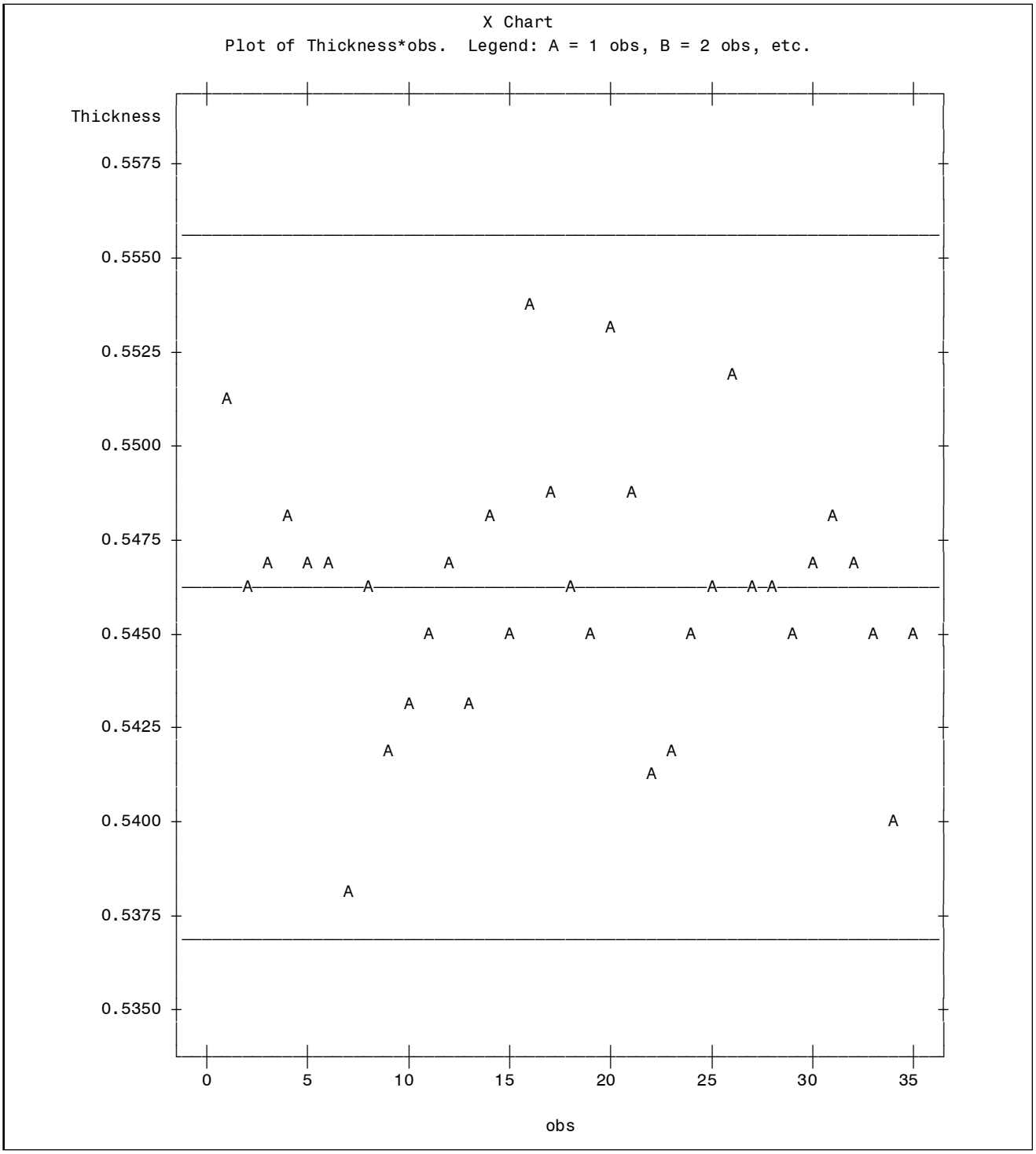
Question: Does plate thickness fall within the process control limits?

SAS Procedure(s): MEANS, PLOT, plus DATA step programming

Code:

```
data ds8;
  input @1 Thickness 5.4;
  obs=_n_;
  * Calculate absolute value of the range pairs. ;
  range = abs(dif(thickness));
datalines;
  .
  .
  .;
run;

proc means data=ds8;
  var thickness range;
  output out=means mean= x_bar r_bar;
run;

data limits;
  set means;
  * Range upper limit ;
  UCL_R = 3.27 * r_bar;
  * Control chart limits ;
  UCL = x_bar + (3*(r_bar/1.128));
  LCL = x_bar - (3*(r_bar/1.128));
  increment_hi = round(UCL,.001) + .001;
  increment_lo = round(LCL,.001) - .001;
run;

data ds9;
  if _n_ = 1 then set limits;
  set ds8;
  * call sumput to create macro variables for use in the plots ;
  call symput('UCL_R',UCL_R);
  call symput('UCL',UCL);
  call symput('LCL',LCL);
  call symput('mean',x_bar);
  call symput('i_hi',increment_hi);
  call symput('i_lo',increment_lo);
run;

proc plot data=ds9;
  plot thickness*obs / vref=&UCL &mean &LCL box;
run;
quit;
```

Output:



Conclusion: Individual thickness values are plotted in relation to the upper control limit (UCL) of .55534 and lower control limit (LCL) of .53688. This control chart indicates that there are no values out of control, i.e., all observations fall within the control limits. There are no shifts or trends present. Therefore, the

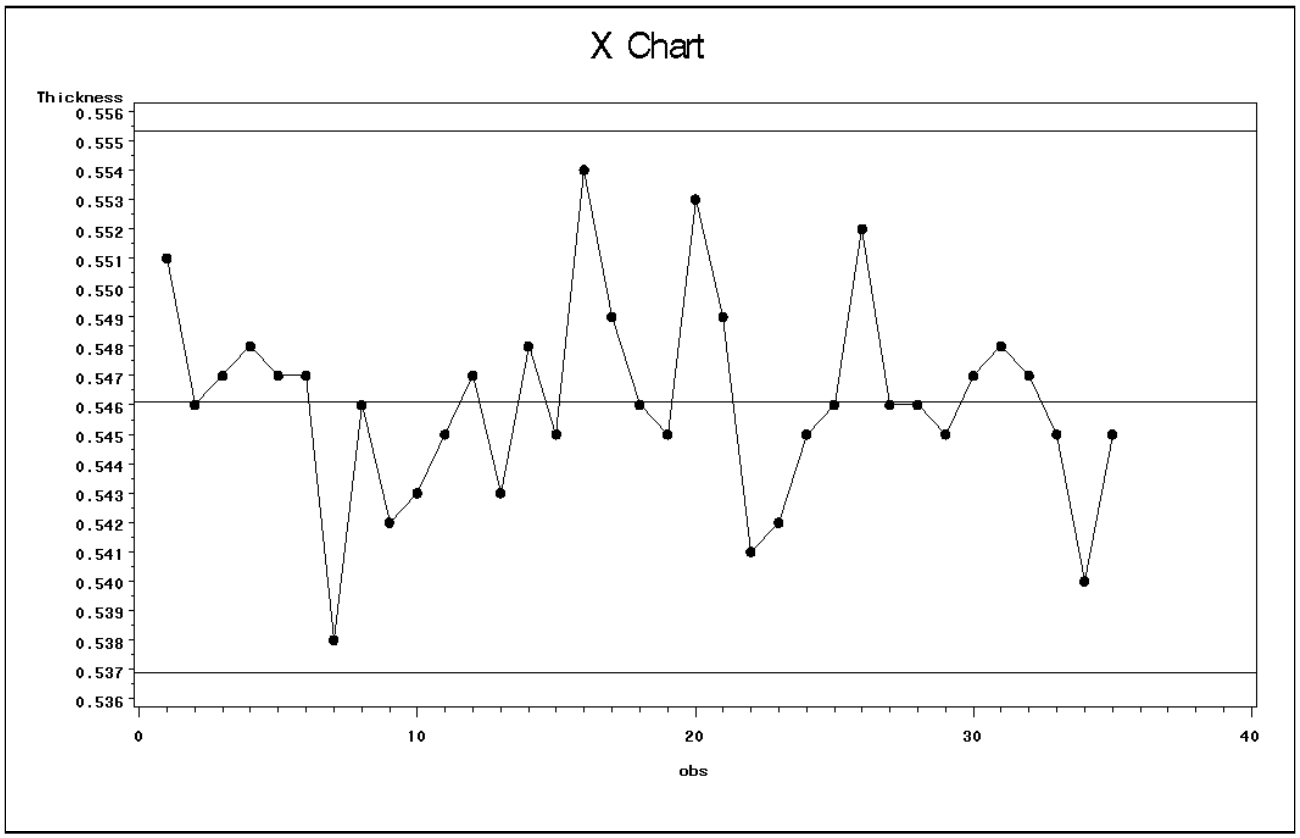
process appears stable and monitoring should continue. In other words, the process is producing product thickness with an acceptable amount of variation.

Optional Approach

Code:

```
symbol interpol=join value=dot;
proc gplot data=ds9;
  plot thickness*obs / vref=&UCL &mean &LCL vaxis = &i_lo to &i_hi by .001;
run;
quit;
```

Enhanced Output:



Issue/Objective #10: The moving range chart displays the absolute value of the difference between sequential pairs of thickness measurements, i.e. the "moving ranges", and compares them to an upper control limit.

Question: Are there any points out of control, i.e., beyond the expected range (the upper control limit)?

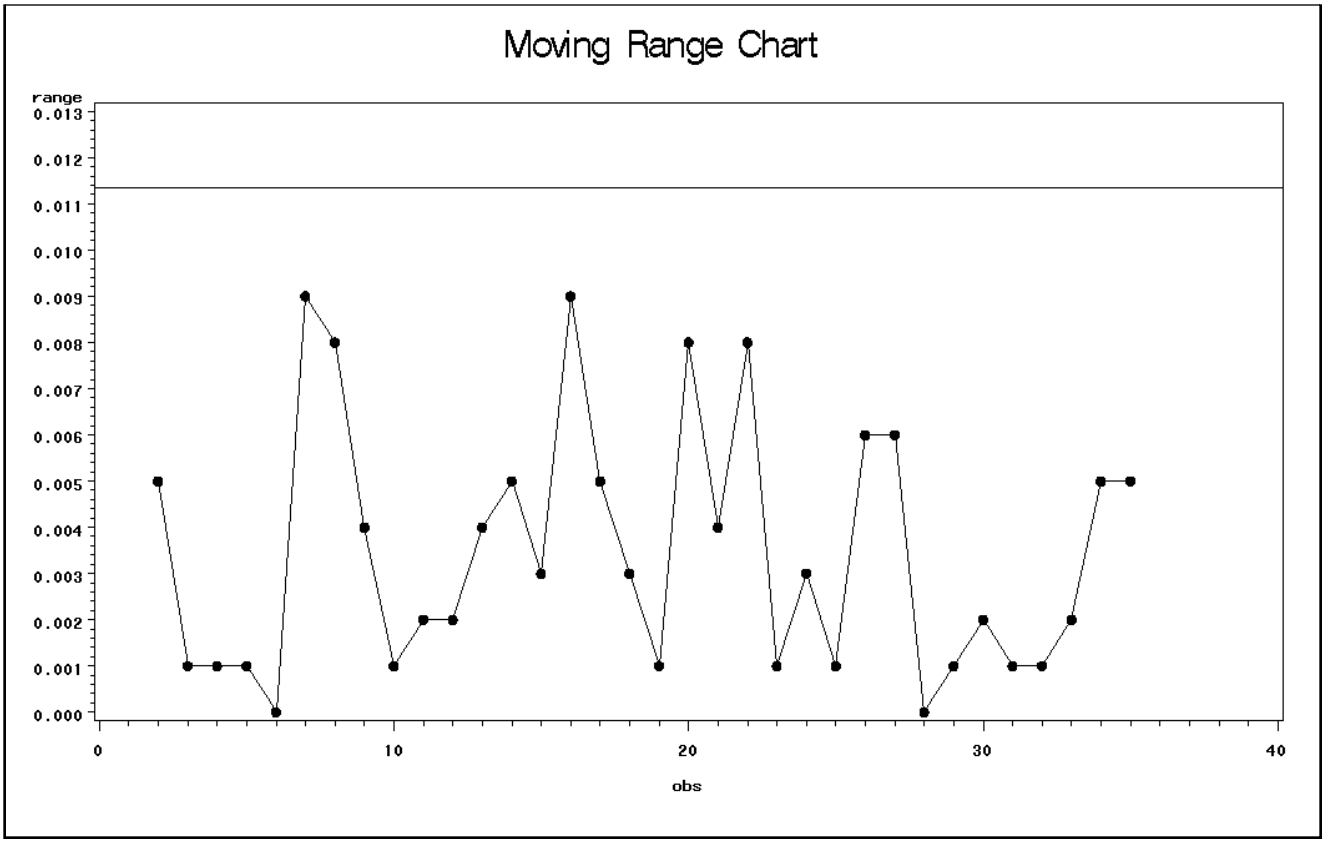
SAS Procedure(s): GPLOT. Note that the upper control limit (UCL_R) for the Moving Range Chart displayed below is calculated in a DATA step on page 25, as UCL_R = 3.27 * r_bar, where:

- r_bar is the average of the absolute difference between each adjacent pair of individual data values for Thickness in the baseline chart data. These difference pairs are referred to as "Moving Ranges".
- 3.27 is a constant obtained from the Control Chart Constants Table for a sample size of n = 2, since the Moving Ranges were calculated using two adjacent data points.

Code:

```
symbol interpol=join value=dot;
proc gplot data=ds9;
  plot range*obs / vref=&UCL_R vaxis = .000 to .013 by .001;
  where range ne .;
run;
quit;
```

Output:



Conclusion: The moving range chart above indicates that measuring in increments of .001 (see data on page 3) provides adequate discrimination, in that the scatter plot of moving range values shows greater than 6 units of measure under the upper control limit of .011349, and there are no points out of control.

Issue/Objective #11: Once a process is deemed stable, process capability can be measured using the Pp and Ppk capability indices, where "capable" can be defined as the likelihood that a stable process will meet the stated specifications and requirements.

Question: Is the process capable of meeting on-going requirements?

SAS Procedure(s): MEANS, plus DATA step programming

Code:

```
*** Mean, Standard Deviation, Pp, Pkp-formula 1, Pkp-formula 2;
proc means data=ds10;
  var thickness;
  output out=stats1 mean= x_bar stddev = sd;
run;

data capability;
  set stats1;
  upper_spec = .56;
  lower_spec = .54;
  Pp = (upper_spec - lower_spec)/(6*sd);
  Pkp_1 = (upper_spec - x_bar)/(3*sd);
  Pkp_2 = (x_bar - lower_spec)/(3*sd);
run;
```

Output:

The MEANS Procedure
Analysis Variable : Thickness

N	Mean	Std Dev	Minimum	Maximum
35	0.5500857	0.0015787	0.5470000	0.5540000

Pp, Ppk

Obs	_FREQ_	x_bar	sd	upper_ spec	lower_ spec	Pp	Pkp_1	Pkp_2
1	35	0.55009	.001578745	0.56	0.54	2.11138	2.09328	2.12948

Conclusion: Pp is the ratio of Specification Spread to Process Spread, where values greater than 1.0 indicate a process that is basically capable of meeting the customer specification, however values greater than 1.5 are desired. In this case, the Pp value is 2.11. Pkp is a performance index which reflects the current process mean's proximity to either the upper specification limit (Pkp_1) or the lower specification limit (Pkp_2), where the smaller of the two values is the selected measure. A value of 1.0 indicates that

the process meets the specification. However, values greater than 1.5 are desired. In this case, the Pkp_1 value is 2.09.

SUMMARY

This is a sample of tasks, not meant to be all inclusive of every statistic or test required for any Six Sigma project. Nor does this paper show the only way to do something. There are almost always multiple ways to complete each of these tasks. The goal of this paper has been to introduce Six Sigma practitioners to the power of SAS as a programming tool for generating output to the analyst, using basic SAS procedures and DATA step programming.

CONTACT INFORMATION

For additional information please contact:

Dan Bretheim
Towers Watson
901 North Glebe Road
Arlington, VA 22203
daniel.bretheim@towerswatson.com