Questions from*:* <http://www.math.wustl.edu/~jmding/math3200/chw/hw2.html>

1. *SAS code under “Problem 1” Heading. See Output pg. 1: The number in the Obs column refers to the sequential baking order as well as the pre-formulated block of dough to be used (i.e. Obs 1 refers to the first worker/method combination which will use the 1st block of pre-formulated dough).*

Initial directions for the pretzel workers A, B, and C: “Each of you will try three different methods of formulating pretzels—Methods 1, 2, and 3—using a different pre-formulated block of dough for each method. With each dough block (in the same oven), you are to make 50 pretzels by the specified method. After using all three methods, you will have each produced 150 pretzels (450 between the three of you). We have randomized both the order of times that the nine batches are baked as well as the method that you are to use. After each baking, we shall announce the next worker by name (A, B, or C) and the method he/she will use (1, 2, or 3). Thank you and let us know if you have any questions! First up is Worker B, Method 3!”

Randomized schedule order (See Output pg. 1):

1st: Worker B, Method 3

2nd: Worker C, Method 2

3rd: Worker C, Method 3

4th: Worker A, Method 1

5th: Worker A, Method 3

6th: Worker C, Method 1

7th: Worker B, Method 2

8th: Worker B, Method 1

9th: Worker A, Method 2

2. *SAS code under “Problem 2” heading*

*(a) See Output pg. 2.* Minimum: 2.86; Lower Quartile (Q1): 6.20**;** Median: 6.405; Upper Quartile (Q3): 6.70; Maximum: 7.00. The summary suggests a left-skewed distribution because Q1 and Q3 are not equidistant from the median (Q2). In fact, the five number summary suggests that while 25% of the data points lie between 2.86 (Min) and 6.20 (Q1) – a range of 3.34 units-- 25% of the data points also lie between 6.70 (Q3) and 7.00 (Max)—a range of only 0.3 units. This suggests a leftward skew. Further, the mean of the data set (6.297) is less than the median. This also gives evidence of a left skewed distribution (as well as the negative value -3.367 under “Skewness”).

*(b)* *See Output pg. 2.* For this data, IQR = Q3-Q1= 6.70-6.20= 0.5. For normally distributed data, IQR is approximately 1.34\*SD. Thus, we can approximate the SD= 0.5/1.34= 0.3731. The sample standard deviation is about 0.788, a larger value than we calculated using IQR; however, we should expect this since the distribution is skewed left. Note: Since the data is skewed, a more detailed breakup of dispersion must be given in terms of the quartiles.

*(c) See Output pg. 3.* The 10% trimmed mean is 6.4335, a higher value than the sample mean of the data (6.297). This suggests that the data contains several outliers on the left side of the distribution (as suggested by the list of extreme observations on Output pg. 4) which matches our expectations for a left-skewed distribution (since the value increased when 10% of the outer data values were trimmed and a data set without outliers would have a trimmed mean approx equal to the untrimmed mean).

*(a2) See Output pg. 4.* The distribution has a left-skewed shape (reflected on the SAS output) with 2 modes (6.70 and 7.00) and 3 outliers (2.86, 2.92, 5.38)

*(b2)* *See Output pg. 3.* Since IQR= 6.70-6.20= 0.5, the lower fence = Q1- 1.5\*IQR= 6.20 -1.5\*0.5= 5.45 and the upper fence = Q3 + 1.5\*IQR= 6.70+1.5\*.5= 7.45. Thus, data points 2.86, 2.92, and 5.38 are all outliers according to the 1.5 x IQR rule.

*(c2) See Output pg. 5*. Given that the shape of the normal plot is reflected when compared to the text, we see that the data appears skewed left with a heavy tailed distribution.

3. *SAS code under “Problem 3” heading.*

*(a) See Output pg. 6.* No, the data does not appear to be normally distributed. Between -1 and +1 SD, the data looks roughly linear; however, there are several outliers and a nonlinear shape outside of 1 standard deviation. In fact, the normal plots maintain a concave shape about the included linear line (below the line for very low or very high scores and above the line towards the middle values). This suggests that the plots might be normalized logarithmically.

(b) *Compare Output pg. 6 and7.* We normalized the data using a logarithmic transformation which provides a weaker shrinking effect of the data values. Looking at the normal plot of log\_score, we see that the transformed data are nearly normally distributed.

4. *SAS code under “Problem 3” heading.*

*(a) See Output pg. 8.* The scatter plot indicates a strong positive relationship (correlation) between average annual snowfall and unemployment percentage. As snowfall increases, unemployment also increases. Conversely, as snowfall decreases, unemployment also appears to decrease.

*(b) See Output pg. 9*. The correlation coefficient between snowfall and unemployment is 0.98352. This can be interpreted as follows: for every additional inch in snowfall, unemployment increases by approximately 0.98352% (and vice versa). This number is significant at the 0.01 level, therefore suggesting a strong linear relationship between snowfall and unemployment.

*(c)* No, this relationship does not mean that the Amherst annual snowfall influences the national unemployment rate or vice versa because a strong correlation does not imply causation. The strong linear relationship between these two variables only means that they are very likely to increase or decrease together (not necessarily because one causes the other).

**PROBLEM 1:**

\*\*For a randomized schedule of the procedures;

OPTIONS LS=**72** PS=**60** CENTER NONUMBER;

**DATA** hw2data1a;

TITLE "HW2 Q1: Randomized Schedule of Pretzel Runs";

INPUT Name$ Method$;

CALL streaminit (**12345**);

RandomValue=rand('uniform');

DATALINES;

WorkerA Method1

WorkerA Method2

WorkerA Method3

WorkerB Method1

WorkerB Method2

WorkerB Method3

WorkerC Method1

WorkerC Method2

WorkerC Method3

;

**proc** **sort** data=hw2data1a OUT=hw2data1b;

by RandomValue;

**RUN**;

**proc** **print** data=hw2data1b (DROP=RandomValue);

**RUN**;

**PROBLEM 2:**

OPTIONS LS=**72** PS=**60** CENTER NONUMBER;

TITLE "HW2 Q2: pH Measurements on Soil Samples";

**data** hw2data2;

INPUT PH @@;

DATALINES;

6.10 6.74 6.22 5.65 6.38 6.70 7.00

6.43 7.00 6.70 6.70 5.94 6.28

6.34 6.62 6.55 2.92 6.10 6.20 6.70

7.00 6.85 6.31 6.26 6.36 6.28

6.38 6.70 6.62 7.00 6.45 6.31

2.86 6.31 6.09 6.17 6.64 6.45

7.00 6.18 6.58 5.38 6.34 7.00

5.70 6.65 6.56 6.00 6.70 6.45

;

\*\*For more readable output of five number summary;

TITLE "HW2 Q2: Five Number Data Summary";

**proc** **means** DATA= hw2data2 min Q1 median Q3 max mean stddev skew ;

Var PH;

**RUN**;

\*\*For more comprehensive summary statistics/ trimmed mean output;

TITLE "HW2 Q2: Trimmed Mean Output";

**proc** **univariate** data=hw2data2 trimmed=**0.1**;

VAR PH;

**RUN**;

\*\*For 4.15, the data is plotted without the 10% trim;

TITLE "HW2 Q2: Plot Outputs";

**proc** **univariate** data=hw2data2 plots;

VAR PH;

PROBPLOT PH/ normal (mu=est sigma=est) pctlminor;

**RUN**;

**PROBLEM 3:**

OPTIONS LS=**72** PS=**60** CENTER NONUMBER;

TITLE "HW2 Q3: Initial Joint Impairment Scores";

**Data** hw2data3;

INPUT Score @@;

log\_score= log(Score);

DATALINES;

6 5 10 9 13 8 8 22 9 18 3

12 11 14 8 32 9 12 17 6 14 9

4 18 18 14 22 9 2 5 3 3 29

;

TITLE "HW2 Q3: Normal Plots";

**proc** **univariate** data=hw2data3;

\*\*Rotated to put Score on X axis to compare with normalized transformation\*

\*Normal Plot of Score;

PROBPLOT Score/normal (mu=est sigma=est) pctlminor ROTATE;

\*Normal Plot of log(Score);

PROBPLOT log\_Score/normal (mu=est sigma=est) pctlminor ROTATE;

**RUN**;

**PROBLEM 4:**

OPTIONS LS=**72** PS=**60** CENTER NONUMBER;

**Data** hw2data4;

TITLE "Annual Snowfall and Unemployment Rate in Amherst, MA";

INPUT Year Snowfall Unemployment@@;

DATALINES;

1973 45 4.9

1974 59 5.6

1975 82 8.5

1976 80 7.7

1977 71 7.1

1978 60 6.1

1979 55 5.8

1980 69 7.1

1981 79 7.6

1982 95 9.7

;

TITLE "Unemployment vs. Snowfall”;

TITLE2 "Scatter Plot";

**PROC** **gPlot** DATA= hw2data4;

PLOT Unemployment\*Snowfall;

**RUN**;

TITLE2 "Covariance and Correlation Coefficient";

**PROC** **corr** DATA=hw2data4 COV;

VAR Unemployment Snowfall;

**RUN**;