Project 1 **Descriptive Statistics** /48

Name _____ **Due 9/12/2014**

I. (10 points)
Problems from *Miller & Freund's Probability and Statistics for Engineers* by R. A. Johnson:

Attach your work and answers as separate sheets. Each solution needs to be organized, neatly written and clearly labeled with its corresponding problem number and attached in the order assigned. If you use Excel on a particular problem, attach a labeled printout of the relevant part of the spreadsheet. **All attached spreadsheet printouts should be set up in Print Preview so that margins, orientation, and scaling yield a readable output that avoids confusing split pages.**

Pages 10 – 11: 1.3, 1.7, 1.8
Pages 22–27: 2.2, 2.25
Pages 36 – 37: 2.34, 2.39, 2.50, 2.51, 2.52

II. (12 points)

Use the following two data sets to compute the requested sums. Put only the answers on this sheet and attach work separately. You may wish to use Excel to do the calculations.

i	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i	3.2	3.5	3.8	4.0	4.3	4.4	4.7	4.8	4.9	5.0	5.1	5.3	5.5
y_i	5.0	5.0	4.9	4.6	4.0	3.9	4.1	3.8	3.7	3.2	3.0	2.8	2.7

Complete the table below and answer the questions that follow. Here the shorthand convention is used that $\sum z_i = \sum_{i=1}^{n} z_i$, n being the number of data points.

Summation Formula	$a = -4.5$	$a = 0$	$a = 4.5$	$a = 9.0$
A. $\sum (x_i - a)$				
B. $\sum x_i - na$				
C. $\frac{\sum (x_i - a)}{n} + a$				
D. $\sum (x_i - a)^2$				
E. $\sum x_i^2 - 2a\sum x_i + na^2$				

What pattern if any do you see between A. and B. above? Prove your assertion.

What do you notice about C.? Prove your assertion.

What pattern if any do you see between D. and E. above? Prove your assertion.

What value of a above makes D. the smallest?

Compute the following:

Summation Formula	**Answer**
$\sum x_i y_i$	_____
$\sum x_i \sum y_i$	_____
$\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$	_____
$\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$	_____
$\sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$	_____
$\frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}}$	_____

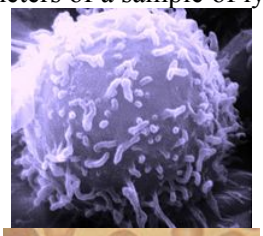
For any fixed set of data x_i, i running from 1 to n, determine the value of a which minimizes each of the following functions.

$f(a) = \left(\sum_{i=1}^{n}\left(x_i - a\right)\right)^2$ minimizing value of a = _____

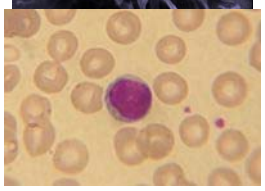
$g(a) = \sum_{i=1}^{n}\left(x_i - a\right)^2$ minimizing value of a = _____

III. (26 points)
The data set that follows are measurements of the average diameters of a sample of lymphocytes.

A scanning electron microscope (SEM) image of a single human lymphocyte (white blood cell).



A stained lymphocyte surrounded by red blood cells viewed using a light microscope.



The measurements are in μm and were extracted from an image analysis of an SEM image. The software which generated the data does not always distinguish between intact lymphocytes and "pieces" or fragments of lymphocytes or between single lymphocytes and adjacent "clusters". This data was supplied by Mike Kostma of MATC's Electron Microscopy program.

16.3975	18.63855	15.20965	18.63918	19.4107	18.38818
3.0105	14.83309	30.56236	20.1218	15.87201	15.17715
15.07998	17.98224	17.05593	44.1056	29.24289	19.45832
19.36242	20.9678	20.82515	18.22499	17.78974	17.33582
18.56787	17.86239	16.55765	21.59801	17.57864	17.29442
18.13202	21.23444	15.53546	13.99835	19.19576	17.07939
17.46741	17.8864	15.85017	21.35174	19.35696	4.472136
2.236068	22.12852	39.16			

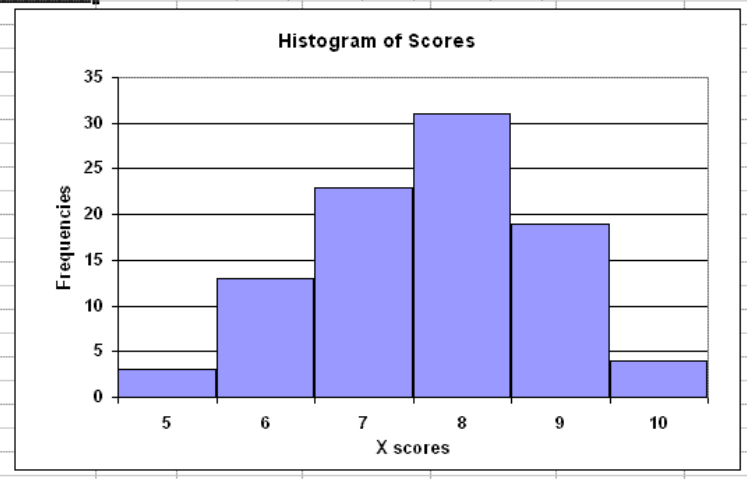
This full data set will be referred to as the 'Ungrouped Data'. For this set of scores calculate and record to the nearest thousandth the descriptive statistics requested in the left side of Table 1 on page 5. Now using a constant class width, group the data so that the lowest class is 1.5-2.5. Using this grouped data, construct a histogram of the scores, plotting either frequency or relative frequency along the vertical axis and the classes along the horizontal axis. Also generate the ogive graph of relative cumulative frequency for less than or equal versus the class boundaries. Use the midpoint or class mark of each class to represent all of the scores in that class. Repeat the same calculations as for the Ungrouped Data and fill in the right side of Table 1, under the heading 'Grouped Data'. Finally, make and attach a box plot of both the Ungrouped and Grouped data.

You may use Excel to present the frequency distributions, do the calculations, and graph the histogram and ogive. Excel does not have a 'built-in' function to calculate the standard deviation of a frequency distribution (the Excel functions STDEV and STDEVP assume each score in the argument list occurs only once.) However, by setting up a column of $f * x$ and a column of $f*x^2$, the standard deviation can be calculated from the formula:

$$s_x = \sqrt{\frac{\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)^2}{n}}{n-1}} \quad ; \quad n = \sum f_i \ .$$

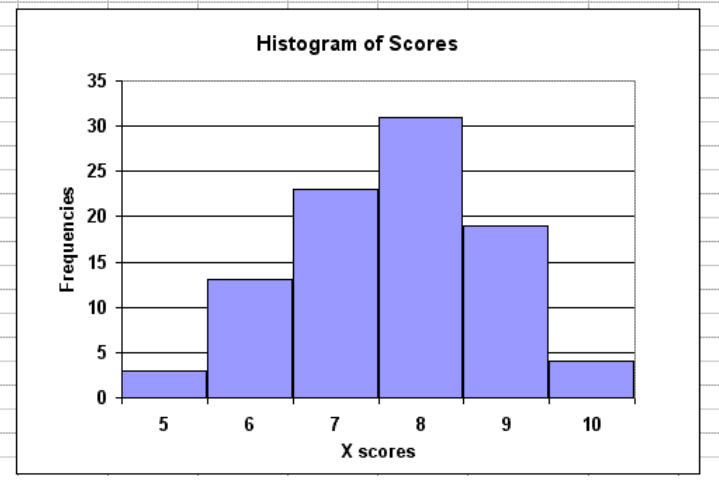
The Excel sample spread sheet shown below illustrates such a calculation.

	A	B	C	D	E	F	G
1	**Formulas Shown in Cells**						
2		**x**	**f**	**x*f**	**x^2*f**		
3		5	3	=B3*C3	=B3^2*C3	**Mean =**	=+D9/C9
4		6	13	=B4*C4	=B4^2*C4	**Sample SD =**	=SQRT((E9-D9^2/C9)/(C9-1))
5		7	23	=B5*C5	=B5^2*C5	**Pop SD=**	=SQRT((E9-D9^2/C9)/(C9))
6		8	31	=B6*C6	=B6^2*C6		
7		9	19	=B7*C7	=B7^2*C7		
8		10	4	=B8*C8	=B8^2*C8		
9			=SUM(C3:C8)	=SUM(D3:D8)	=SUM(E3:E8)		



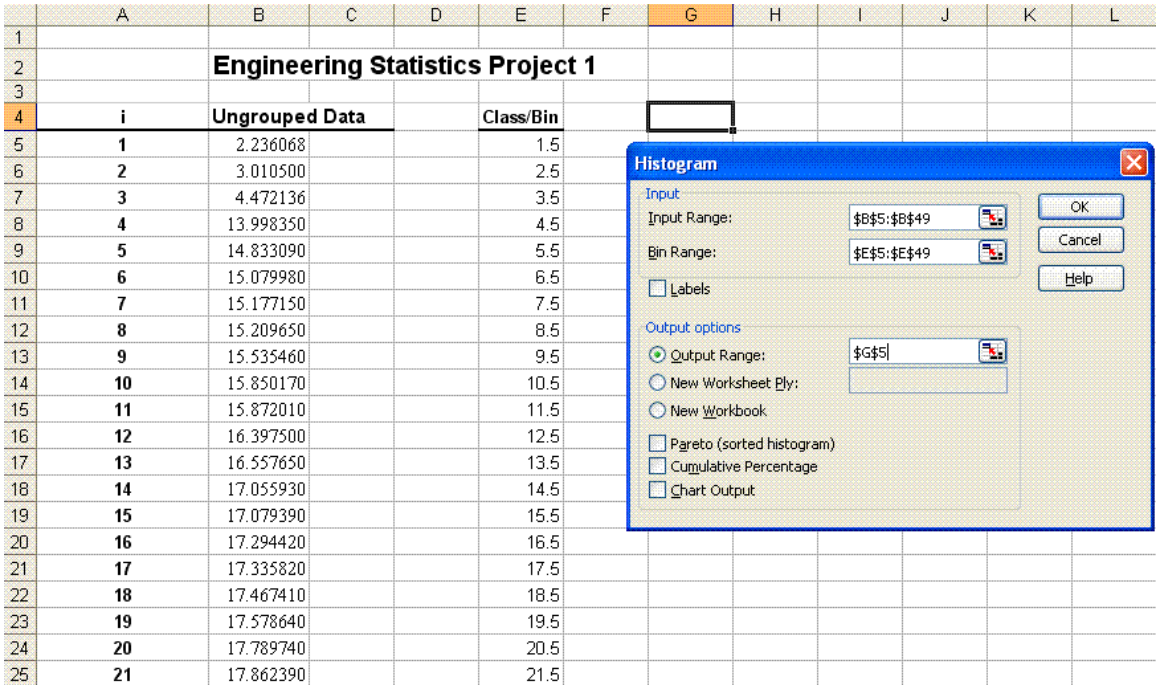
The output of the above formulas is shown below.

	A	B	C	D	E	F	G	H	I
1									
2		**x**	**f**	**x*f**	**x^2*f**				
3		5	3	15	75		**Mean =**	7.666667	
4		6	13	78	468		**Sample SD =**	1.173376	
5		7	23	161	1127		**Pop SD=**	1.167051	
6		8	31	248	1984				
7		9	19	171	1539				
8		10	4	40	400				
9			93	713	5593				



In Excel to have the histogram bars fill up the class width as shown above, click on one of the rectangles in the histogram, then right click and select Format Data Series from the right-click menu. In the Format Data series menu select Series Options and set the Gap Width to 0%. To generate the ogive graph in Excel choose a chart type that is a line graph of connected points.

Excel can even generate the frequency distribution of the classes. This requires the Data Analysis package be available under the Tools menu. If Data Analysis is not shown in the Data Menu, click on the Office Button and choose Excel Options, then choose Add-Ins. From the list of Add Ins available: check Anaylsis ToolPak and click **Go**. You will need to setup a column of left-class boundaries for the grouped data. Excel calls the column of these boundaries a "Bin Range". Once the Data Analysis Tool is chosen from the Tools menu, select Histogram and click OK. From the Histogram menu select the Input Range as the cells in the column of the ungrouped data, the Bin Range as the column of **Right** class boundaries, and then pick a cell where you want the resulting frequency distribution to begin as the Output Range. Click OK to generate the frequency distribution.

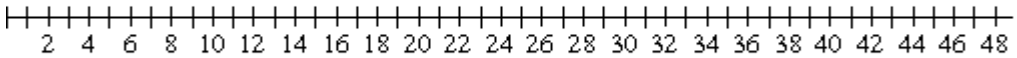


	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		**Engineering Statistics Project 1**										
3												
4	**i**	**Ungrouped Data**			**Class/Bin**							
5	**1**	2.236068			1.5							
6	**2**	3.010500			2.5							
7	**3**	4.472136			3.5							
8	**4**	13.998350			4.5							
9	**5**	14.833090			5.5							
10	**6**	15.079980			6.5							
11	**7**	15.177150			7.5							
12	**8**	15.209650			8.5							
13	**9**	15.535460			9.5							
14	**10**	15.850170			10.5							
15	**11**	15.872010			11.5							
16	**12**	16.397500			12.5							
17	**13**	16.557650			13.5							
18	**14**	17.055930			14.5							
19	**15**	17.079390			15.5							
20	**16**	17.294420			16.5							
21	**17**	17.335820			17.5							
22	**18**	17.467410			18.5							
23	**19**	17.578640			19.5							
24	**20**	17.789740			20.5							
25	**21**	17.862390			21.5							

Table 1

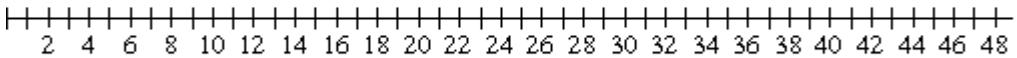
Descriptive Statistic	Ungrouped Data	Grouped Data
Minimum		
Maximum		
Range		
Mode	**N. A.**	
Median, M_d		
Mean, $\bar{x}$		
Q_1		
Q_3		
IQR		
60'th Percentile, P_{60}		
Sample Standard Deviation, s_x		
Population Standard Deviation, σ_x		
Sample coefficient of variation		
Sample Variance, s_x^2		
Population Variance, σ_x^2		

Box Plot of Ungrouped Data



Diameter in microns

Box Plot of Grouped Data



Diameter in microns

How closely do the descriptive statistics of the grouped and ungrouped scores compare?

How well does grouping the scores into classes represent the actual data?

For the ungrouped data, what fraction of the scores is within one standard deviation of the mean?

For the ungrouped data, what fraction of the scores is within two standard deviations of the mean?

For the ungrouped data, what fraction of the scores is within three standard deviations of the mean?

For the ungrouped data, what fraction of the scores have a z score larger than 1 ?

For the ungrouped data, what fraction of the scores have a z score smaller than 1 ?

For the ungrouped data, what fraction of the scores have a $|z|$ (absolute value of z score) larger than 1 ?

Using the box plot of the ungrouped data, eliminate all "outliers", i.e., all data beyond the outer fence (3.0 IQR's from the box hinges). Now recompute the mean and the sample standard deviation of this reduced data set.

Sample Mean $\overline{x}$ = _____

Sample Standard Deviation s_x = _____

Eliminating the outliers changed the values of both of these statistics. Compared to the results for the full ungrouped data set, which statistic showed the greatest percent change by eliminating the outliers? Explain this observation.

In general, if outliers are eliminated, will the value of this same statistic increase or decrease? Explain your answer.

For this data set give a possible justification for eliminating the outliers.

Project 2 **Basic Probability** /48

Name _____ **Due 9/22/2014**

I. (13 points)
Problems from *Miller & Freund's Probability and Statistics for Engineers* by R. A. Johnson:

Attach your work and answers as separate sheets. Each solution needs to be organized, neatly written and clearly labeled with its corresponding problem number and attached in the order assigned. If you use Excel on a particular problem, attach a labeled printout of the relevant part of the spreadsheet. **All attached spreadsheet printouts should be set up in Print Preview so that margins, orientation, and scaling yield a readable output that avoids confusing split pages.**

Pages 54 – 65: 3.13, 3.17, 3.25, 3.26
Pages 64–66: 3.34, 3.41, 3.42, 3.48, 3.51
Pages 75 – 76: 3.58, 3.61, 3.67, 3.68

II.
1. (6 points)
A test is performed on a manufactured product to detect a specific electronic defect. On a product **without** this defect the test will erroneously indicate the presence of the defect 0.1% of the time. On a product **with** this defect the test will erroneously fail to detect it 1.0% of the time. Past history indicates that 4.25% of the products have this electronic defect. Calculate the following:

a) What percent of all products don't have this specific defect?

b) What percent of all products don't have this specific defect and the test confirms this condition?

c) What percent of all products don't have this specific defect but the test says otherwise?

d) What percent of all products have this specific defect and test fails to detect it?

e) What percent of all products have this specific defect and test confirms this condition?

f) For what percent of all products does the test give incorrect results?

g) Given that a product tests as having this specific defect, what is the probability that it really does not have the defect?

h) Given that a product tests as having this specific defect, what is the probability that it really does have the defect?

i) Given that a product tests as not having this specific defect, what is the probability that it really does not have the defect?

j) Given that a product tests as not having this specific defect, what is the probability that it really does have the defect?

2. (2 points)
A student must answer 10 of 13 questions (each worth 10 points) on a 100 point exam.
a) How many different choices as to which set of questions to answer does any one student have?

b) Answer the same question assuming that the exam rules state that everyone must answer the first five questions.

3. (5 points)
Four integrated circuits (IC's) are to be sampled for testing from a lot of 39 experimental prototypes produced.
a) How many different samples are possible?

b) What is the probability that a particular group of 4 IC's out of the 39 produced would be the ones chosen for testing?

c) Suppose that five of the 39 IC's are defective. Let x be the number of defective IC's chosen in the sample of four to be tested. Fill in the following probability distribution:

x	$p(x)$
0	
1	
2	
3	
4	

4. (4 points)
a) A manufacturing process consists of four sub assembly operations performed in series. If each step is 99% reliable, what is the overall reliability of the process?

b) A power supply is rated at 95% reliability and it has two separate backups each rated as 70% reliable. Assuming that the failure of a power supply is independent of the other power supplies, what is the probability of having power?

5. (5 points)
Suppose that on any given flight of a particular kind of aircraft that the chance of an aileron malfunction is 0.015%. Assume that this probability never changes and that having an aileron malfunction is a random process.
a) Calculate the probability that an aircraft of this kind has an aileron malfunction on its first flight.

b) Calculate the probability that an aircraft of this kind will have an aileron malfunction on its 1000'th flight, given that it has already completed 999 flights without incident.

c) Calculate the probability that an aircraft of this kind makes 999 flights without incident and then has an aileron malfunction on its 1000'thflight.

d) Calculate the probability that an aircraft of this kind makes 1000 flights and never has an aileron malfunction.

e) Calculate the probability that an aircraft of this kind has an aileron malfunction sometime before it completes its 1000'th flight.

III. (13 points)
Toss two six-sided dice 144 times. For each toss record the sum of the two uppermost faces. From this data construct the relative frequency distribution (i.e., the empirical probability distribution) to the nearest ten-thousandth. Using the assumption of a fair experiment (i.e., the "classical probability concept"), calculate the theoretical probabilities, also to the nearest ten-thousandth. The Mean value of the probability distribution is the mathematical expectation (expectation value) of **the sum of faces**. The population variance is the expectation value of the squared deviation of **the sum of faces** from its mean value. The population standard deviation is the square root of the population variance. For the 144 die tosses, the mean and population standard deviation are just the mean and population standard deviation of your 144 scores.

x = Sum of Faces	Empirical Probability	Theoretical Probability
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
Mean Value of x		
Population Standard Deviation of x		

Using Excel show both the empirical and the theoretical probability distributions as probability histograms. Use a series legend to identify each distribution. Use the classes 1.5-2.5, 2.5-3.5, etc., so that each "score" (possible sum value) is associated with its own class of width of one. On the vertical axis plot the respective probability (empirical or theoretical). Attach a print out of your histogram with your assignment.

How do the empirical and theoretical distributions differ?

Why are the two distributions different?

Assuming a "fair toss" of the dice, what would you need to do to make the two distributions converge?

What is the area to the right of 5.5 under the theoretical probability distribution curve?

Interpret this area as a probability.

How well do the means and standard deviations of the two distributions compare?

Project 3 **Discrete Probability Distributions** /48

Name _____ **Due 9/29/2014**

I. (11 points)
Problems from *Miller & Freund's Probability and Statistics for Engineers* by R. A. Johnson:

Attach your work and answers as separate sheets. Each solution needs to be organized, neatly written and clearly labeled with its corresponding problem number and attached in the order assigned. If you use Excel on a particular problem, attach a labeled printout of the relevant part of the spreadsheet. **All attached spreadsheet printouts should be set up in Print Preview so that margins, orientation, and scaling yield a readable output that avoids confusing split pages.**

Pages 91 – 93	4.3, 4.5, 4.6, 4.7, 4.8, 4.12, 4.20, 4.25
Pages 102–103:	4.34, 4.41, 4.44, 4.47, 4.48
Pages 110 – 111:	4.50, 4.51, 4.56, 4.59, 4.61, 4.67, 4.68, 4.69
Page 112:	4.74

II.
1. (7 points) A state LOTTO game is advertised as follows:

Win Up To $3,000,000! Play Number-Buck!

Rules: For only $1 purchase any six numbers of your choice from 1 to 49. In the grand drawing six different numbers from 1 to 49 are picked at random. You compare your numbers to those drawn. Depending on the number of matches you win as follows:

3 matches	$2
4 matches	$30
5 matches	$400
6 matches	$3,000,000

Your friend, Joe Hapless, who, sadly, never benefited from a course in statistics, is impressed with the ad and wants to play. As his more enlightened confidant you agree to compute the probabilities of his winning. Let the random variable, x, stand for the number of matches.

a) Let $f(x)$ be the probability that the number of matches is x and fill in the probability distribution table for x. Calculate the mean and the standard deviation of this distribution and record them as well.

b) Let m stand for Joe's net earnings and let $q(m)$ be the probability that Joe earns m. Remember, Joe must pay $1 just to play the game! Complete the probability distribution table for m and also calculate the values for the mean and standard deviation of m.

Probability Distribution Table for *x*

x	***f*(*x*)**
0	
1	
2	
3	
4	
5	
6	
$\mu_x = \langle x \rangle$	
$\sigma_x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$	

Probability Distribution Table for *m*

M	***x***	***q*(*m*)**
−1.00	$0 \le x \le 2$	
	3	
	4	
	5	
	6	
$\mu_m = \langle m \rangle$		
$\sigma_m = \sqrt{\langle m^2 \rangle - \langle m \rangle^2}$		

c) If 1,000,000 tickets for this lotto were sold, how much money would you expect the game's sponsors (i.e., the state) to make?

d) On the average how many tickets for this game must be sold for there to be a jackpot ($3,000,000) winner?

e) How many tickets for this game must be sold so that the probability of at least one jackpot winner becomes 95%?

f) For the number of tickets sold in part e) how many jackpot winners would you expect?

g) For the number of tickets sold in part e) how much money would you expect the state to make?

2. (4 points) The discrete function $f(x,y)$ has the values shown in the following table and is zero for all other inputs. Let $f_1(x)$ be the marginal pdf of x and $f_2(y)$ be the marginal pdf of y.

	x				
y	**0**	**1**	**2**	**3**	**4**
−1	0.050	0.100	0.150	0.150	0.050
$xyf(x,y)$					
$f_1(x)f_2(y)$					
0	0.015	0.075	0.105	0.090	0.015
$xyf(x,y)$					
$f_1(x)f_2(y)$					
1	0.010	0.070	0.050	0.050	0.020
$xyf(x,y)$					
$f_1(x)f_2(y)$					

a) If x and y are interpreted as discrete random variables, could this function represent their joint probability density function? Explain.

b) Fill in the following:

X	$f_1(x)$	$xf_1(x)$	$x^2 f_1(x)$
0			
1			
2			
3			
4			
Column Sum			

Y	$f_2(y)$	$yf_2(y)$	$y^2 f_2(y)$
-1			
0			
1			
Column Sum			

c) Compute the following:

$\mu_x =$
$\sigma_x =$
$\mu_y =$
$\sigma_y =$
$\langle xy \rangle =$
$\text{cov}(x,y) =$
$\rho(x,y) =$

d) Are x and y independent random variables? Explain.

3. (4 points) The discrete function $f(x,y)$ has the values shown in the following table and is zero for all other inputs. Let $f_1(x)$ be the marginal pdf of x and $f_2(y)$ be the marginal pdf of y.

	x				
y	**0**	**1**	**2**	**3**	**4**
−1	0.025	0.050	0.100	0.050	0.025
$xyf(x,y)$					
$f_1(x)f_2(y)$					
0	0.050	0.100	0.200	0.100	0.050
$xyf(x,y)$					
$f_1(x)f_2(y)$					
1	0.025	0.050	0.100	0.050	0.025
$xyf(x,y)$					
$f_1(x)f_2(y)$					

a) If x and y are interpreted as discrete random variables, could this function represent their joint probability density function? Explain.

b) Fill in the following:

X	$f_1(x)$	$xf_1(x)$	$x^2 f_1(x)$
0			
1			
2			
3			
4			
Column Sum			

Y	$f_2(y)$	$yf_2(y)$	$y^2 f_2(y)$
-1			
0			
1			
Column Sum			

c) Compute the following:

$\mu_x =$
$\sigma_x =$
$\mu_y =$
$\sigma_y =$
$\langle xy \rangle =$
$\text{cov}(x,y) =$
$\rho(x,y) =$

d) Are x and y independent random variables? Explain.

4. (4 points) The discrete function $f(x,y)$ has the values shown in the following table and is zero for all other inputs. Let $f_1(x)$ be the marginal pdf of x and $f_2(y)$ be the marginal pdf of y.

	x				
y	**0**	**1**	**2**	**3**	**4**
−1	0.025	0.100	0.050	0.050	0.100
$xyf(x,y)$					
$f_1(x)f_2(y)$					
0	0.050	0.050	0.050	0.100	0.100
$xyf(x,y)$					
$f_1(x)f_2(y)$					
1	0.085	0.035	0.020	0.065	0.120
$xyf(x,y)$					
$f_1(x)f_2(y)$					

a) If x and y are interpreted as discrete random variables, could this function represent their joint probability density function? Explain.

b) Fill in the following:

X	$f_1(x)$	$xf_1(x)$	$x^2 f_1(x)$
0			
1			
2			
3			
4			
Column Sum			

Y	$f_2(y)$	$yf_2(y)$	$y^2 f_2(y)$
-1			
0			
1			
Column Sum			

c) Compute the following:

$\mu_x =$
$\sigma_x =$
$\mu_y =$
$\sigma_y =$
$\langle xy \rangle =$
$\text{cov}(x,y) =$
$\rho(x,y) =$

d) Are x and y independent random variables? Explain.

5. (2 points) Calculate the following to three significant digits (for part b to achieve the stated accuracy you might want to use Wolfram Alpha).

a) 80!

b) 80,000,000!

III.

1. (8 points)

a) In a group of n people, none of whom were born on a leap day (February 29 of a leap year), what is the probability that at least two people share a birthday? Assume that peoples' birthdays are uniformly distributed over the 365 days of the year.

b) In Excel generate a column in a spreadsheet that calculates this probability from $n = 1$ to $n = 100$. [**Hint:** Let $q(n)$ be the probability that in the group of n people no common birthdays occur. Develop a recursion for $q(n)$ in terms of n and $q(n-1)$. Using relative addressing this recursion is easy to implement in Excel.] Also insert in your spreadsheet a scatter plot which displays the probability of at least one common birthday versus n for $n = 1$ to $n = 100$.

c) What is the first value of n for which the probability of at least one common birthday exceeds 50%?

d) **(2 point Bonus)** Using Stirling's formula, determine a formula for $G(n)$, so that $1-e^{G(n)}$ approximates the probability of at least one common birthday in the group of n people. As a check, in your spreadsheet add another column adjacent to the column of part b) which displays the results of this approximation $n = 1$ to $n = 100$.

Print out and attach a copy of your spreadsheet with your assignment.

2. (8 points)

Toss ten coins 100 times and for each of the 100 tosses record the number of heads, x. From the resulting frequency distribution of x, compute the empirical probability (relative frequency) of each value of x. Compute the theoretical probabilities assuming fair and independent tosses of the ten coins. **Note:** A single experiment in this coin toss scenario is defined as a toss of the ten coins.

There are then 100 repetitions of this same experiment. Calculate the mean and standard deviations of both the theoretical and empirical distributions and enter all results in the table below.

Using Excel superimpose both the empirical and the theoretical probability distributions as probability histograms. Use a series legend to identify each distribution. Use classes -0.5-0.5, 0.5-1.5, 1.5-2.5, etc., so that each "score" (possible sum value) is associated with its own class of width of one. On the vertical axis plot the respective probability (empirical or theoretical). Attach a print out of your histogram with your assignment.

X	Empirical Probability	Theoretical Probability
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
$\mu_x = \langle x \rangle$		
$\sigma_x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$		

How do the empirical and theoretical distributions differ?

Why are the two distributions different?

What is the theoretical probability of 7 or more heads on one toss of the ten coins?

What is the theoretical probability of 4 or less heads on one toss of the ten coins?

Do the means of the theoretical and empirical distributions match better or worse than most of the specific event probabilities?

Give a reason why this makes sense.

Project 4 **Continuous Probability Distributions** /48

Name _____ **Due 10/13/2014**

I. (12 points)
Problems from *Miller & Freund's Probability and Statistics for Engineers* by R. A. Johnson:

Attach your work and answers as separate sheets. Each solution needs to be organized, neatly written and clearly labeled with its corresponding problem number and attached in the order assigned. If you use Excel on a particular problem, attach a labeled printout of the relevant part of the spreadsheet. **All attached spreadsheet printouts should be set up in Print Preview so that margins, orientation, and scaling yield a readable output that avoids confusing split pages.**

Pages 124 – 125:	5.2, 5.3, 5.13
Pages 133 – 134:	5.20, 5.21, 5.22, 5.24, 5.28
Pages 144 – 145:	5.45, 5.58, 5.61, 5.62
Pages 156 – 157:	5.72, 5.73, 5.74, 5.75, 5.85, 5.90, 5.91
Page 163:	5.94, 5.96, 5.97, 5.99, 5.100

II. Each Problem below is worth 5 points.
For Problems 1 and 2 you are given the distribution (cumulative probability distribution) function $F(x)$ which give the probability that the value of the random variable is less than or equal to x. From this determine

i) the probability density function $f(x)$
ii) the first quartile, Q_1 , the median M_d , and the third quartile Q_3
iii) the mean or expected or expectation value of x: $\mu_x = E(x) = \langle x \rangle$
iv) the expected or expectation value of x^2: $E\left(x^2\right) = \left\langle x^2 \right\rangle$
v) σ_x

1. For $a > 0$, $F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 1 - e^{-x/a} & \text{if } x > 0 \end{cases}$

$f(x) =$

$Q_1 =$ _____ $M_d =$ _____

$Q_3 =$ _____ $\mu_x = \langle x \rangle =$ _____

$\left\langle x^2 \right\rangle =$ _____ $\sigma_x =$ _____

2. For $a > 0$, $F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 - \cos\left(\dfrac{\pi x}{2a}\right) & \text{if} \quad 0 \le x < a \\ 1 & \text{if} \quad x > a \end{cases}$

$f(x) =$

$Q_1 =$ _____ $M_d =$ _____

$Q_3 =$ _____ $\mu_x = \langle x \rangle =$ _____

$\langle x^2 \rangle =$ _____ $\sigma_x =$ _____

3. The probability density function of the standard normal distribution is given by $f(z) = Ae^{-z^2/2}$. The constant A is called a **normalization factor** and is determined by the requirement that $\int_{-\infty}^{\infty} f(z)dz = 1$. The formula for A can be discovered by the following "trick".

$$B = \int_{-\infty}^{\infty} e^{-z^2/2}dz = \sqrt{\left(\int_{-\infty}^{\infty} e^{-x^2/2}dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}dy\right)} = \sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\left(x^2+y^2\right)/2}dydx}$$

a) Transform the double integral to polar coordinates and from the formula for B determine A.

b) For positive σ determine the normalization factor A in the probability density function $f(x) = Ae^{-(x-\mu)^2/(2\sigma^2)}$.

c) Calculate $E\left(\left[x-\mu\right]^2\right) = \left\langle (x-\mu)^2 \right\rangle$ for the probability distribution of b).

For problems 4 and 5 the techniques illustrated in the online notes at
http://faculty.matcmadison.edu/alehnen/EngineeringStats/BivariateNormalDistribution.pdf
may proof helpful.

4. For positive a and L, the joint probability density function for continuous random variables x and y is given by the following function.

$$f(x,y)=\begin{cases} 0 & \text{if } x<0 \\ \dfrac{e^{-(y-b)^2/(2a^2)}}{\sqrt{2\pi}La}=\dfrac{\exp\left(-(y-b)^2/\left(2a^2\right)\right)}{\sqrt{2\pi}La} & \text{if } 0\le x\le L \\ 0 & \text{if } x>L \end{cases}$$

b) Let $f_1(x)$ be the marginal pdf of x and $f_2(y)$ be the marginal pdf of y.

$f_1(x)=$

$\langle x\rangle=$ _____ $\mu_x=$ _____

$\langle x^2\rangle=$ _____ $\sigma_x=$ _____

$f_2(y)=$

$\langle y\rangle=$ _____ $\mu_y=$ _____

$\langle y^2\rangle=$ _____ $\sigma_y=$ _____

c) Compute the following:

$\langle xy\rangle=$
$\operatorname{cov}(x,y)=$
$\rho(x,y)=$

d) Are x and y independent random variables? Explain.

5. (**5 point bonus**) For positive *a* and *b*, the joint probability density function for continuous random variables *x* and *y* is given by the following function.

$$f(x,y)=\begin{cases} 0 & \text{if } x<0 \text{ or } y<0 \\ \dfrac{4(x+y)e^{-\left[(x/a)^2+(y/b)^2\right]}}{\sqrt{\pi}ab(a+b)} & \text{if } x\geq 0 \text{ and } y\geq 0 \end{cases}$$

b) Let $f_1(x)$ be the marginal pdf of *x* and $f_2(y)$ be the marginal pdf of *y*.

$f_1(x) =$

$\langle x \rangle =$ _____ $\mu_x =$ _____

$\langle x^2 \rangle =$ _____ $\sigma_x =$ _____

$f_2(y) =$

$\langle y \rangle =$ _____ $\mu_y =$ _____

$\langle y^2 \rangle =$ _____ $\sigma_y =$ _____

c) Compute the following:

$\langle xy \rangle =$

$\text{cov}(x,y) =$

$\rho(x,y) =$

d) Are *x* and *y* independent random variables? Explain.

III.

1. (10 points)

The probability that a standard normal variable, ξ, takes on a value between 0 and Z is given by the integral $\Pr(0<\xi<Z)=G(z)=\frac{1}{\sqrt{2\pi}}\int_0^Z e^{-t^2/2}dt$.

a) Use the Maclaurin series for e^x to generate a power series for $\Pr(0<\xi<Z)$.

b) What is the interval of convergence of the power series in a)

c) If $\Pr(0<\xi<Z)=G(z)=\frac{1}{\sqrt{2\pi}}\sum_{j=0}^{\infty}a_j(Z)$, where $a_j(Z)$ are non-zero terms in the power series of part a), determine the following:

$a_0(Z)=$_____ $\frac{a_j(Z)}{a_{j-1}(Z)}=$_____

d) Using your results from c), generate an Excel spreadsheet with the following three labeled columns.

First: Z running from 0 to 5 in steps of 0.1

Second: Excel's approximation to $\Pr(0<\xi<Z)$ formatted to eight decimal places. Use the built in standard normal probability distribution function, NORMSDIST( ). Since this function gives the cumulative probability that $\Pr(\xi<Z)$ you will need to subtract 0.5 from its output to get $\Pr(0<\xi<Z)$.

Third: The result of summing $\frac{1}{\sqrt{2\pi}}\sum_{j=0}^{50}a_j(Z)$ also formatted to eight decimal places. To do this sum, for each row that goes with a given value of Z, generate 51 columns, one for each term you're going to sum. Use the recursion implied by your answer to part c) to generate the terms.

	$\Pr(0<\xi<Z)$	$\Pr(0<\xi<Z)$	***j***	**0**	**1**
Z	**NORMSDIST(Z)-0.5**	**Power Series**	***j*'thTerm label →**		
0.10	0.03982784	0.03982784		1.0000E-01	-1.6667E-04
0.20	0.07925971	0.07925971		2.0000E-01	-1.3333E-03
0.30	0.11791142	0.11791142		3.0000E-01	-4.5000E-03

e) Comment on the accuracy of Excel's built in function compared to the power series approximation for the values of *Z* considered.

f) Add one additional row to the bottom with $Z = 9$. What happens?! Try to come up with an explanation for what you see.

g) Print out your spreadsheet but only select the first three labeled columns shown above. Attach your printout with the assignment.

2. (3 points)
Now consider the probability that a standard normal variable, ξ, takes on a value greater than *Z*. $\Pr(\xi > Z) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Z} e^{-t^2/2}dt = \frac{1}{\sqrt{2\pi}}\int_{Z}^{\infty} e^{-t^2/2}dt$. As the last part of Problem 6 illustrated getting accurate answers as *Z* gets large can be a problem. An alternate approach is an asymptotic approximation to the probability. If you make the substitution $u = t^2/2$ and integrate by parts,

$$\Pr(\xi > Z) = \frac{1}{\sqrt{2\pi}}\int_{Z^2/2}^{\infty}\frac{e^{-u}}{\sqrt{2u}}du = \frac{1}{2\sqrt{\pi}}\left[\left(\frac{-e^{-u}}{\sqrt{u}}\right)\right]_{Z^2/2}^{\infty} - \frac{1}{2}\int_{Z^2/2}^{\infty}\frac{e^{-u}}{u^{3/2}}du\Bigg]$$

$$= \frac{e^{-Z^2/2}}{\sqrt{2\pi}Z} - \frac{1}{4\sqrt{\pi}}\int_{Z^2/2}^{\infty}\frac{e^{-u}}{u^{3/2}}du$$

Now, the second integral is of order $\frac{e^{-Z^2/2}}{Z^{3/2}}$, which is $\sqrt{Z}$ times smaller than the first term. So we have the asymptotic representation that as $Z \to \infty$, $\Pr(\xi > Z) \to \frac{e^{-Z^2/2}}{\sqrt{2\pi}Z}$. Generate an Excel spreadsheet with the following **three** labeled columns.

a) Z running from 1 to 25 in steps of 1

b) Excel's approximation to $\Pr(\xi > Z)$ formatted as scientific with four decimal places. Use the built in standard normal probability distribution function, NORMSDIST( ). Since this function gives the cumulative probability that $\Pr(\xi < Z)$ you will need to subtract it from 1 to get $\Pr(\xi > Z)$.

c) The asymptotic approximation to $\Pr(\xi > Z)$ also formatted as Scientific with four decimal places.

To help you check your work the actual probabilities to five significant figures are given below for the first eight values of Z.

Z	$\Pr(\xi > Z)$
1	1.5866E-01
2	2.2750E-02
3	1.3499E-03
4	3.1671E-05
5	2.8665E-07
6	9.8659E-10
7	1.2798E-12
8	6.2210E-16

Print out and attach your spread sheet with the assignment.

Comment on how well the asymptotic formula represents $\Pr(\xi > Z)$.

3. (3 points)
Let p be the probability that a score in a standard normal distribution is less than z, i.e.

$$p = F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt .$$

Since $F(x)$ is an increasing function, an inverse function exists with $z = F^{-1}(p)$. In Excel this inverse is called NORMSINV() . For a given p, z can be calculated by numerical methods.

a) Determine $L(x)$ the linear approximation to $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$ about the origin.

$L(x) =$

b) Solve for z_0 in the equation $L(z_0) = p$.

$z_0 =$

c) $F^{-1}(p)$ is the only real root of the equation $H(z) = F(z) - p = 0$. Determine an explicit formula for the Newton's method iteration for solution of this equation.

$$z_{n+1} = z_n - \frac{H(z_n)}{H'(z_n)}.$$

d) Use a numerical integration procedure on your calculator or a computer to calculate the Newton's method iterations up to z_5 or until successive iterations differ in absolute value by less than 10^{-8}. Use z_0 from b) as the initial guess. Fill in the following table stating all z values to nine decimal places.

	p			
	0.1	**0.2**	**0.6**	**0.7**
z_0				
z_1				
z_2				
z_3				
z_4				
z_5				
Excel's NORMSINV(***p***)				

Project 5 **Sampling Distributions** /48
Name _____ **Due 10/29/2014**

I. (23 points)
Problems from *Miller & Freund's Probability and Statistics for Engineers* by R. A. Johnson:

Attach your work and answers as separate sheets. Each solution needs to be organized, neatly written and clearly labeled with its corresponding problem number and attached in the order assigned. If you use Excel on a particular problem, attach a labeled printout of the relevant part of the spreadsheet. **All attached spreadsheet printouts should be set up in Print Preview so that margins, orientation, and scaling yield a readable output that avoids confusing split pages.**

Pages 186 – 187:	6.5, 6.9, 6.11, 6.12, 6.13, 6.14, 6.18
Pages 191 – 192:	6.20, 6.22, 6.23, 6.25, 6.27, 6.28
Page 194:	6.30, 6.33
Page 196:	6.34, 6.36, 6.37
Page 201:	6.42, 6.43, 6.45, 6.46, 6.48

II. (3 points)
a) A standard six sided die is tossed fairly and the number of dots on the top face, x, is recorded. Let μ_1 and σ_1^2 be the mean and variance respectively of x. Fill in the following:

$\mu_1 =$

$\sigma_1^2 =$

$\sigma_1 =$

b) Suppose now that n distinguishable six sided dice are tossed fairly. The random variable x is defined as the sum of the dots on the resulting top faces. Let μ_x and σ_x^2 be the mean and variance respectively of x. Fill in the following:

Range of x =

$\mu_x =$

$\sigma_x^2 =$

$\sigma_x =$

c) A six sided die is tossed 420 times and the number of dots on the top face is recorded for each toss. Let x = sum of these recorded numbers. If all of the tosses were fair calculate the following probabilities.

$P(1420 \le x \le 1520) =$	_____	$P(x > 1500) =$	_____
$P(x < 1400) =$	_____	$P(x = 1425) =$	_____
$P(x = 1420) =$	_____	$P(1420 \le x \le 1425) =$	_____

III.
1. (3 points)
Use the set of 45 average lymphocyte diameters was given in Project 1.

a) Generate an Excel spreadsheet that makes a normal scores plot of the data. First enter the lymphocyte diameters as a column of 45 entries and then sort this column in ascending order. Next generate a parallel column of integers, *j*, from 1 to 45. The Excel Inverse Standard Normal Distribution Function, NORMSINV(P), returns the *z* score which has an area to the left equal to P. Parallel to the other two columns column to enter a third column generated by NORMSINV(j/46). This will provide a column of 45 *z* scores distributed across N(0, 1) with equal probability between each score. Make a scatter plot of the lymphocyte diameters versus the z scores.

b) Based on the normal scores plot, how normal is the distribution of lymphocyte diameters?

c) Now, as in Project1, remove all "outliers", i.e., all data beyond the outer fence (3.0 IQR's from the box hinges). Then repeat the normal scores analysis on this reduced data set.

d) Based on the normal scores plot, how normal is the distribution of lymphocyte diameters once the outliers have been removed?

e) Hand in your spreadsheet with the assignment.

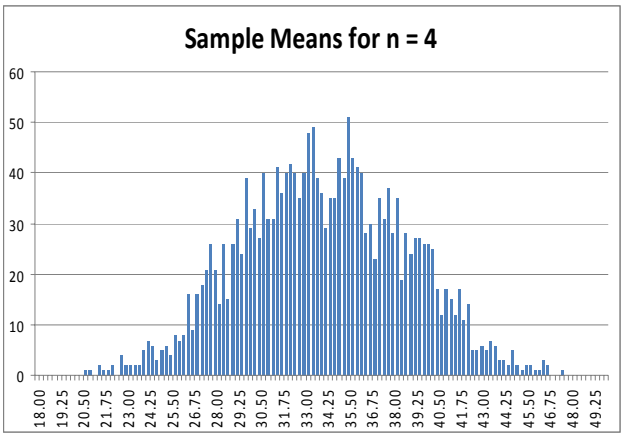
2. (11 points)
In Excel we are going to simulate the sampling distribution of means and the sampling distribution of variances for scores drawn from a uniform distribution. You will approximate the sampling distributions of these statistics by generating 2000 random samples of given sample size (first $n = 4$ and then $n = 36$) from a uniform parent population with a mean of 34, a minimum of 18, and a maximum of 50. To do this, generate an index column of integers from 1 to 2000. Start this column in the middle of the spread sheet so as to leave ample rows near the top for summary results and histograms of the sampling distributions.

Next to this index column for the sample generate four columns of data with each column consisting of 2000 random scores drawn from a uniform distribution on [18, 50]. To generate a score from this parent population use the formula 32*RAND()+18. Each row of these four columns represents a random sample of size 4 drawn from the parent population. The random number generator RAND() will recompute these scores every time **any** new formula is entered or computed in the spreadsheet. To prevent the generated samples from constantly changing, select and copy all four of these data columns. Position the cursor to the first cell of the first data column and from the Edit menu, choose Paste Special. In this menu check "Values" and then click OK. Your 2000 samples of sample size 4 are now "fixed". Now add two additional columns to the right of the last data column. In each row of the first additional column compute the sample average of the four randomly drawn scores in this row. In each row of the second additional column compute the sample variance of the four randomly drawn scores in this row.

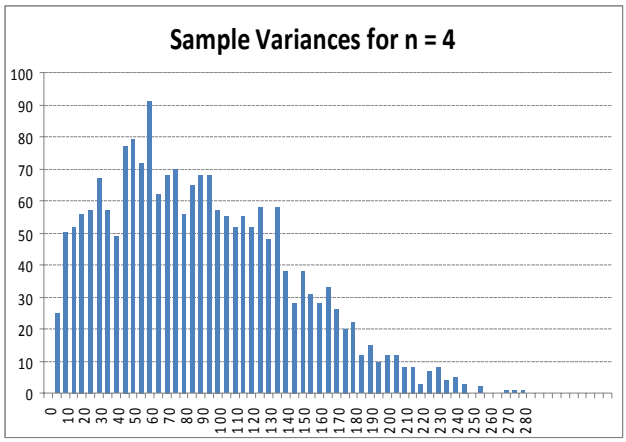
Compute the mean, minimum, maximum, and standard deviation of the 2000 sample means and the mean, minimum, maximum, and standard deviation of the 2000 sample variances. Display the results in labeled cells near the top of the spreadsheet. Include labeled cells for the values of the mean, variance and standard deviation of the parent population as well as the standard error of the mean for samples of size $n = 4$. Create a column of right class boundaries for the sample means that goes from the minimum sample mean to the maximum sample mean in steps of 0.25. Leaving several blank columns, create a second column of right class boundaries for the sample variances that goes from the minimum sample variance to the maximum sample variance in steps of 5.

Now from the Data menu select Data Analysis, from the menu select Histogram. For the Input Range select the column of 2000 sample averages. Select the column of right class boundaries for the sample means as the Bin Range and choose a convenient cell as the Output Range where the grouped frequency distribution of sample means will begin in the spreadsheet. From this grouped frequency distribution construct a histogram of the 2000 sample averages. Select Histogram from the Data Analysis menu. For the Input Range select the column of 2000 sample variances. Select the column of right class boundaries for the sample variances as the Bin Range and choose a convenient cell as the Output Range where the grouped frequency distribution of sample variances will begin in the spreadsheet. From this grouped frequency distribution construct a histogram of the 2000 sample variances. Place both histograms near the top of the spreadsheet by the summary data for samples of size $n = 4$. Examples of the histograms are displayed below.

Distribution of Sample Means



Distribution of Sample Variances



Now generate a second set of 2000 samples, only this time use a sample size of 36. This means 36 columns of data again followed by a column of the sample means and a column of the sample variances. Repeat all of the steps you carried out for the samples of size 4. For both histograms you should use a reasonable value for the step size in the column of right class boundaries that is smaller than the values used when $n = 4$. Place the new two histograms of the sample statistics based on $n = 36$ as well as labeled values of the mean, standard deviation, minimum and maximum of each distribution at the top of the spreadsheet under the previous results for the sampling distributions based on $n = 4$. Set Print Area to include only that part of the spreadsheet that contains the four histograms and the summary statistics. After adjusting the margins and size in Setup to get a readable printout that avoids confusing split pages, print your selection. Fill in the table below and answer the following questions.

Summary Table of Sampling Distributions from the Uniform Population

Parent Population Mean μ_x	
Parent Pop Variance σ_x^2	
Parent Pop Standard Deviation σ_x	
***n* = 4: Mean of the 2000 sample means**	
***n* = 4: Standard Deviation of the 2000 sample means**	
***n* = 4: Standard Error of the Mean** $\sigma_{\bar{x}}$	
***n* = 4: Mean of the 2000 sample variances**	
***n* = 4: Standard deviation of the 2000 sample variances**	
***n* = 4:** $\sqrt{2/\upsilon}\,\sigma_x^2$	
***n* = 36: Mean of the 2000 sample means**	
***n* =36: Standard Deviation of the 2000 sample means**	
***n* = 36: Standard Error of the Mean** $\sigma_{\bar{x}}$	
***n* = 36: Mean of the 2000 sample variances**	
***n* = 36: Standard deviation of the 2000 sample variances**	
***n* = 36:** $\sqrt{2/\upsilon}\,\sigma_x^2$	

a) For the samples of size 4, how closely does the mean of the 2000 sample means match the population mean?

b) For the samples of size 4, how does the spread of the sampling distribution of sample means compare to the spread of the parent population?

c) For the samples of size 4, how does the shape of the sampling distribution of sample means compare to the shape of the parent population?

d) For the samples of size 4, how closely does the standard deviation of the 2000 sample means match the standard error of the mean?

e) For the samples of size 4, how closely does the mean of the 2000 sample variances match the variance of the parent population?

f) For the samples of size 4, how closely does the standard deviation of the 2000 variances match $\sqrt{2/\upsilon}\,\sigma_x^2$?

g) For the samples of size 36, how closely does the mean of the 2000 sample means match the population mean?

h) For the samples of size 36, how does the spread of the sampling distribution of sample means compare to the spread of the sampling distribution of sample means based on samples of size 4?

i) For the samples of size 36, how does the shape of the sampling distribution of sample means compare to the shape of the sampling distribution of sample means based on samples of size 4?

j) For the samples of size 36, how closely does the standard deviation of the 2000 sample means match the standard error of the mean?

k) For the samples of size 36, how closely does the mean of the 2000 sample variances match the variance of the parent population?

l) For the samples of size 36, how closely does the standard deviation of the 2000 sample variances match $\sqrt{2/\upsilon}\,\sigma_x^2$?

m) Based on this simulation, which parameter, the mean or variance, is better estimated by its sample statistic? Explain your answer.

3. (8 points)
Go to the website http://onlinestatbook.com/stat_sim/sampling_dist/index.html to run a Java simulation of sampling distributions from various parent populations. In the instructions it states:

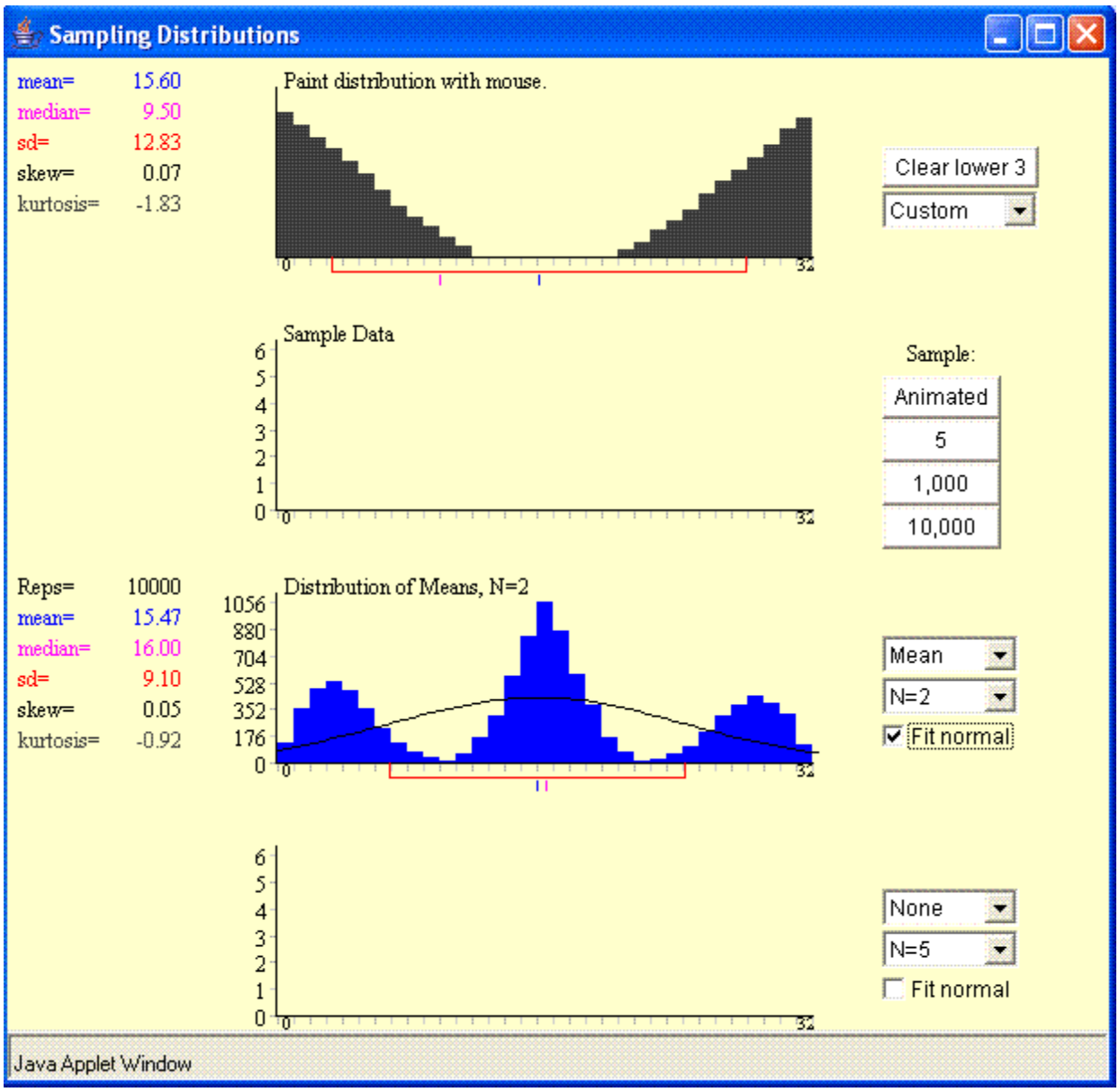
"Please wait until a button appears below". The button is **Actually to the Left!**. Click the Begin button to start the simulation. There will be a slight delay and then the Java applet opens in a new window. When the applet begins, a histogram of a normal distribution is displayed at the top of the screen.

The distribution portrayed at the top of the screen is the population from which samples are taken. The mean of the distribution is indicated by a small blue line and the median is indicated by a small purple line. When the mean and median are the same, the two lines overlap. The red line extends from the mean one standard deviation in each direction.

The second histogram displays the sample data. This histogram is initially blank. The third and fourth histograms show the distribution of statistics computed from the sample data. The number of samples (replications) that the third and fourth histograms are based on is indicated by the label "Reps=." This number is chosen from the three options stated, 5, 1000, or 10,000. Choose **10,000** ten times so that the total sample size, i.e. Reps = 100,000. The "Animated" button samples 5 scores and illustrates the random sampling. Two sampling distributions can be displayed and for each one you can specify which statistic and sample size to use with the pop-up menu. However, **leave the bottom one blank**. This will avoid scaling problems, since by default the same scale is used on both distributions. The scale required for the variance, which is larger than the scale needed for the mean, would also be applied to the mean if both distributions are generated in parallel. On the first sampling distribution choose **Mean** to display the distribution of sample means and later choose **Var(U)** to display the distribution of the **sample** variances. **Note:** you need to choose **Var(U)** option, not the Variance option. Variance gives a biased estimate of the population variance by dividing the sum of squared deviations by n instead of $n-1$. The **Var(U)** divides the sum of squared deviations by $n-1$ so that it's estimate of the population variance is unbiased. The sample size of each sample can be set to 2, 5, 10, 16, 20 or 25 from the pop-up menu. Do **not** confuse the **sample size** with the **number of samples** (which is supposed to be 100,000 if you press the **10,000** button ten times).

Numerical values of the statistics of the sampling distribution are displayed to the left of the histogram. By clicking the "Fit normal" button you can see a normal distribution superimposed over the simulated sampling distribution. The pop-up menu at the top of the screen allows you to change the parent population. There are four options for the parent population: Normal, Uniform, Skewed and Custom.

In your simulation you will use three different parent populations each to be sampled with $n = 2$, $n = 16$ and $n = 25$. The populations to use are Normal, Uniform and BiModal. You will need to construct the BiModal population. From the top pop-up menu choose Custom. By clicking on the top histogram with the mouse and dragging, generate a "BiModal" distribution similar to the one shown below.



For each parent population and each sample size simulate both the distribution of sample means and the distribution of sample variances. As **noted above**, do them **one at a time**. This avoids compressing the shape of the sampling distribution of means which results if both distributions are generated in parallel. Fill in the two tables below using the results displayed by the simulation.

The Sampling Distribution of Means

Population	Normal	Uniform	BiModal
Parent Population Mean μ_x			
Parent Pop Standard Deviation σ_x			
***n* = 2: Mean of the Sampling Distribution**			
***n* = 2: S. D. of the Sampling Distribution**			
***n* = 2: Standard Error of the Mean** $\sigma_{\bar{x}}$			
***n* = 16: Mean of the Sampling Distribution**			
***n* = 16: S. D. of the Sampling Distribution**			
***n* = 16: Standard Error of the Mean** $\sigma_{\bar{x}}$			
***n* = 25: Mean of the Sampling Distribution**			
***n* = 25: S. D. of the Sampling Distribution**			
***n* = 25: Standard Error of the Mean** $\sigma_{\bar{x}}$			

Table 2: The Sampling Distribution of Variances

Population	Normal	Uniform	BiModal
Parent Pop Variance σ_x^2			
***n* = 2: Mean of the Sampling Distribution**			
***n* = 2: S. D. of the Sampling Distribution**			
***n* = 2:** $\sqrt{2/\upsilon}\,\sigma_x^2$			
***n* = 16: Mean of the Sampling Distribution**			
***n* = 16: S. D. of the Sampling Distribution**			
***n* = 16:** $\sqrt{2/\upsilon}\,\sigma_x^2$			
***n* = 25: Mean of the Sampling Distribution**			
***n* = 25: S. D. of the Sampling Distribution**			
***n* = 25:** $\sqrt{2/\upsilon}\,\sigma_x^2$			

a) For which population did the $n = 2$ sampling distribution of means least resemble a normal distribution? Explain if this is surprising or not.

b) For every parent population what happened to the width of the sampling distributions of means as *n* increased? In what way was this reflected in the numbers you recorded in Table 1?

c) For every parent population what happened to the shape of the sampling distribution of means as *n* increased?

d) In general, how did the values of the standard deviations of the sampling distribution of means compare with the standard error of the mean?

e) How are the results of this simulation consistent with the Central Limit Theorem?

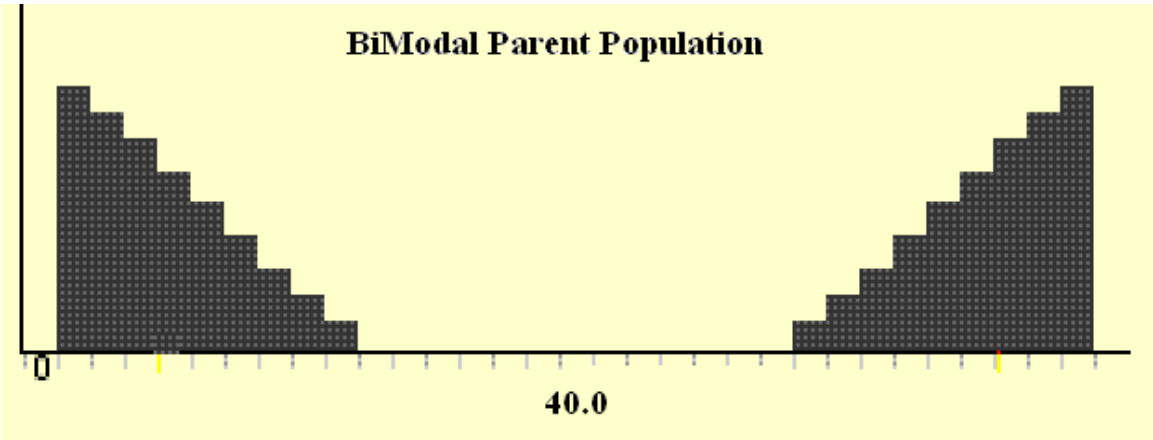
f) How well do the means of the sampling distribution of variances approximate the population variance? Is this surprising?

g) For every parent population what happened to the shape of the sampling distribution of variances as *n* increased?

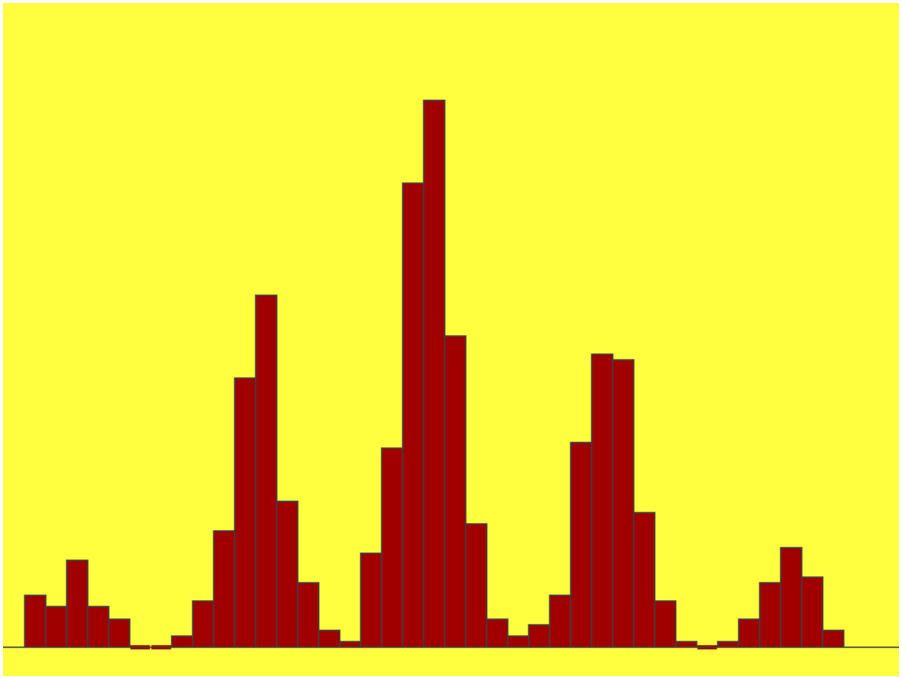
h) For which parent population did the values of the standard deviation of the sampling distribution of variances come closest to $\sqrt{2/\upsilon}\,\sigma_x^2$? Explain why this makes sense.

i) Under what condition is the sampling distribution of $(n-1)\left(\frac{s_x}{\sigma_x}\right)^2$ a χ^2 distribution with $n-1$ degrees of freedom? Based on this simulation, how important is this condition for the result to be true?

j) The graph below shows a parent population similar to the "BiModal" distribution used in the Sampling Distributions simulation. The parent population parameters were $\mu = 40$ and $\sigma = 17.01777$.



Below is shown a detailed view of the histogram of the sampling distribution of means for 750 samples each of sample size *n* taken at random from this population. Based on this histogram determine the sample size *n* and estimate the mean and standard deviation of these 750 sample averages.



Project 6 **Statistical Decision Theory** /48

Name _____ **Due 11/4/2014**

I. (34 points)
Problems from *Miller & Freund's Probability and Statistics for Engineers* by R. A. Johnson:

Attach your work and answers as separate sheets. Each solution needs to be organized, neatly written and clearly labeled with its corresponding problem number and attached in the order assigned. If you use Excel on a particular problem, attach a labeled printout of the relevant part of the spreadsheet. **All attached spreadsheet printouts should be set up in Print Preview so that margins, orientation, and scaling yield a readable output that avoids confusing split pages.**

Pages 212 – 214:	7.1, 7.6, 7.20, 7.22
Page 228:	7.40, 7.42, 7.43, 7.46, 7.49
Pages 234 – 235:	7.53, 7.58, 7.62
Page 236:	7.68
Page 241:	7.71
Page 258:	8.3, 8.6, 8.10
Page 262:	8.14, 8.15
Page 264:	8.23
Page 270:	9.3, 9.4
Pages 275 – 276:	9.7, 9.15
Pages 282 – 283:	10.1, 10.2, 10.9, 10.10
Pages 290 – 291:	10.19, 10.25, 10.27, 10.29, 10.30, 10.31

II.
1. (5 points)
At a R&D center there are five different research groups. Some research groups are perceived both internally and externally as being more prestigious than others. The center maintains its own private library, but the library's share of the total research budget is insufficient to purchase all the materials requested at any given time. Therefore, when a member of a particular research group makes a request, he or she must also specify a priority, from 1 (the highest) to 4 (the lowest), which indicates how urgently the item is needed. The following contingency table summarizes all requests for the first six months of the last fiscal year.

Research Group → / **Priority ↓**	**A**	**B**	**C**	**D**	**E**	**Row Sum**
1	46	33	48	36	33	
2	39	30	41	30	26	
3	19	13	23	26	16	
4	16	7	18	26	14	
Column Sum						

Test, at a 5% level of significance whether there is a relationship between the assignment of priorities and the research group the request comes from. State and explain your conclusion. Compute and report the *P*- Value of your observed statistic.

2. (4 points)
In Project 2, Part III, you tossed two six-sided dice 144 times and recorded the observed frequencies for the sum of the two faces. Based on your data is there evidence at a 5% level of significance that the die-toss was not done fairly? To answer this question fill in the table below.

Sum of Faces	Theoretical Probability	Expected f	Observed f	$(O-E)^2/E$
2 or 3				
4				
5				
6				
7				
8				
9				
10				
11 or 12				

State and explain your conclusion. Compute and report the *P*- Value of your observed statistic.

3. (5 points)
In Excel use the same set of average lymphocyte diameters as in Project 1. Set up the Table in Excel with the data grouped in the six classes shown in the table below.

Class Boundaries							
Lower Bound	**Upper Bound**	**f**	***z* lower**	***z* upper**	**Norm Pr**	***E***	$(O-E)^2/E$
–∞	**13.5**						
13.5	**15.5**						
15.5	**17.5**						
17.5	**19.5**						
19.5	**21.5**						
21.5	∞						
Column Sum							

For calculation purposes you need to use the mean and (sample) standard deviation for the ungrouped data. For each class calculate the *z* scores of the lower and upper class boundaries (called *z* lower and *z* upper, respectively). In Excel use the function NORMSDIST() to calculate the area between these two *z* scores in the standard normal distribution, N(0,1) . This is the probability of the given class assuming that the scores are normally distributed. From this model probability you can calculate an expected frequency to compare against observation. Generate a histogram which displays both the expected and observed frequencies versus the classes.

Is there evidence at a 5% level of significance that the lymphocyte diameters are not normally distributed? State and explain your conclusion. Explain how you reached your conclusion. Compute and report the *P*- Value of your observed statistic.

Now, as in Project 1, remove all “outliers”, i.e., all data beyond the outer fence (3.0 IQR's from the box hinges). Then group the scores into at least five classes in such a way that no expected frequency is less than 5. For this reduced data set generate a histogram which displays both the expected and observed frequencies versus the classes. Is there evidence at a 5% level of significance that the lymphocyte diameters with outliers removed are not normally distributed? Explain how you reached your conclusion. Compute and report the *P*- Value of your observed statistic.

Hand in your spreadsheet with the assignment.

Project 7 **One Way Analysis of Variance and Regression Analysis** /48

Name _____ **Due 12/11/2014**

I. (12 points)
Problems from *Miller & Freund's Probability and Statistics for Engineers* by R. A. Johnson:

Attach your work and answers as separate sheets. Each solution needs to be organized, neatly written and clearly labeled with its corresponding problem number and attached in the order assigned. If you use Excel on a particular problem, attach a labeled printout of the relevant part of the spreadsheet. **All attached spreadsheet printouts should be set up in Print Preview so that margins, orientation, and scaling yield a readable output that avoids confusing split pages.**

Pages 368 – 369:	12.2, 12.10
Pages 316 – 318:	11.4, 11.5, 11.20
Page 331:	11.26
Page 344:	11.50, 11.51

II.
1. (16 points)
A pilot study on three-member wood joints held with a single steel bolt was performed to determine if previous load conditions affected the strength of the joints. All joints were made to the same design from Douglas-fir.

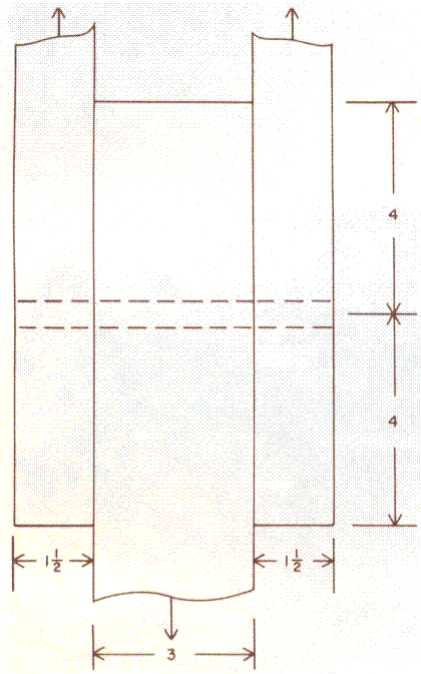


Figure 1—Specimen geometry. Member thickness was 1-1/2 inches. (All dimensions are inches.) (ML88 5387)

The specimens were three-member wood joints (fig. 1) with a single bolt. The main members were 3 inches wide and the side members were 1-1/2 inches wide. All members were 1-1/2 inches thick. All material was Douglas-fir with a moisture content of approximately 10 percent.
Bolts were 1/2 inch in diameter and 7 inches in length. They conformed to ASTM Standard A 307 (Low Carbon Steel Externally and Internally Threaded Standard Fasteners). Washers were used under the bolt head and the nut. Nuts were finger-tight. The bolt holes were 9/16 inch in diameter.

The specimens had a ratio of main member width-to-bolt diameter of six (Mode II failure). This ratio was selected because it is commonly found in glulam construction.

End distance was 4 inches in both main and side members.

The following four load conditions were used:

A: A short duration ramp load (i.e., a control group)
B: A constant load of 4080 lbs applied for 1 year
C: A constant load of 2880 lbs applied for 1 year
D: A constant load of 1440 lbs applied for 1 year

The strengths of the joints were then determined in destructive testing by monitoring the maximum load (pounds) to joint failure. The data is shown in the table below.

Previous Load Condition →	**A**	**B**	**C**	**D**
	4420	5950	5140	4810
	4570	4670	5030	5370
	4600	5830	4870	5480
	4640	4870	4920	5540
	4440	4510	5560	6580
	4860	5590	5110	6200
	4620		5090	7410
	5020			5810

In Excel set up a one-way analysis of variance as is described in the ANOVA notes at http://faculty.matcmadison.edu/alehnen/EngineeringStats/ANOVA.pdf . Since the numbers are rather large, you will have display problems. One possible solution is to format results in scientific notation. A second solution is to subtract a constant from every score as this will shift the means, but not the variances. The Excel function FDIST gives the probability in a F distribution with the appropriate observed F score and numerator and denominator degrees of freedom. The Excel function FINV gives critical one-tail F scores for a stated level of significance and the appropriate numerator and denominator degrees of freedom.

Test at a 5% level of significance whether any of the load condition variations are associated with differences in maximum load. State and explain your conclusion. Explain how you reached your conclusion. Compute and report the *P*- Value of your observed statistic.

Now determine by a Bonferroni procedure which, if any, of the four load conditions are significantly different from each other in their mean maximum load. Remember that the Excel function TINV outputs the two tail critical degree score.

What assumptions are necessary for the above Single Factor ANOVA to be valid?

Do these assumptions seem to be satisfied by this particular set of data? Explain your answer.

Hand in your spreadsheet with the assignment.

2. (20 points)
Measurements of the electrical resistivity of tungsten at specified temperatures are presented in the table below. The error in the determination of the temperature was negligible. The temperatures are stated as absolute temperatures in °K. The units on resistivity are $\mu\Omega$cm .

Temperature	Resistivity	Temperature	Resistivity
300	5.65	1200	30.98
400	8.06	1300	34.08
500	10.56	1400	37.19
600	13.23	1500	40.36
700	16.09	1600	43.55
800	19.00	1700	46.78
900	21.94	1800	49.69
1000	24.93	1900	53.35
1100	27.94	2000	56.67

In Excel perform a regression analysis as detailed in the Regression Notes found at http://faculty.matcmadison.edu/alehnen/EngineeringStats/regression.pdf . Your spreadsheet analysis of this data set should contain two scatter plots. The first should display both the data and the predictions of the linear regression model drawn as a continuous line. The second should display the residuals plotted versus the control variable. Assuming that the underlying relation between absolute temperature and resistivity for tungsten is linear and that the errors in the measurements are normally distributed, calculate the lower and upper limits for a 99% confidence intervals requested in the following table.

99% Confidence Interval	Lower Limit	Upper Limit
Population Slope β_1		
Population Intercept β_0		
The resistivity when T =1150°K		
A measurement of resistivity when T =1150°K		
The resistivity when T =1950°K		
A measurement of resistivity when T =1950°K		

Why are the width of the confidence intervals for T =1150°K and T =1950° K different?

Fill in the following Regression ANOVA table as shown below.

Source	Sum of Squares	Degrees of Freedom	Mean Square
Linear Model			
Error			
Total			

Observed F score _____

At a 1% level of significance, what conclusion do you draw from this ANOVA Table?

Comment on how well the linear model fits this data.

What does the pattern of residuals indicate about the adequacy of the linear model?

One of the data points is in fact incorrect. Which one is it? (Justify your answer.)

The resistivity measurement for the data point in question is in fact too low. Correct it by adding 0.36 $\mu\Omega$cm to the value stated in the table. In the spreadsheet add a second linear regression analysis on this corrected data set. Be sure to include the scatter plots of the data plus regression line as well as the scatter plot of the residuals. Fill in the following table.

	Original Data Set	Corrected Data Set
Regression Slope		
Regression Intercept		
Correlation Coefficient		
Coefficient of Determination		
Standard Error of Estimate		

What conclusion can you make about using r^2 as the sole measure of the accuracy of a model?

The quadratic regression model (it's still linear in the parameters) for fitting a set of data, (x_i, y_i), $1 \le i \le n$, to a parabola is given by $\hat{y} = \hat{\beta}_{2,0} + \hat{\beta}_{2,1}x + \hat{\beta}_{2,2}x^2$. The values of $\hat{\beta}_{2,0}$, $\hat{\beta}_{2,1}$ and $\hat{\beta}_{2,2}$ are determined by minimizing the error variation given by

$$\sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_{2,0} + \hat{\beta}_{2,1}x_i + \hat{\beta}_{2,2}x_i^2 \right) \right]^2 .$$

Write out and solve the normal equations for the quadratic regression. Attach your work on separate sheets. The process of solving and the form of the final equations is "cleaner" if you use the following notation.

Let $C(u,v) = \sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v}) = \sum_{i=1}^{n}u_i(v_i - \overline{v}) = \sum_{i=1}^{n}v_i(u_i - \overline{u}) = \sum_{i=1}^{n}u_i v_i - \dfrac{\left(\sum_{i=1}^{n}u_i\right)\left(\sum_{i=1}^{n}v_i\right)}{n}$.

$\overline{x} = \dfrac{\sum_{i=1}^{n}x_i}{n}$ $\quad \overline{y} = \dfrac{\sum_{i=1}^{n}y_i}{n}$ $\quad \overline{x^2} = \dfrac{\sum_{i=1}^{n}x_i^2}{n}$ $\quad C(x,x) = SS_{xx}$

$C(x,y) = SS_{xy}$ $\quad C\left(x^2,x\right) = \sum_{i=1}^{n}x_i^2(x_i - \overline{x})$ $\quad C\left(x^2,x^2\right) = \sum_{i=1}^{n}\left(x_i^2 - \overline{x^2}\right)^2$ $\quad C\left(x^2,y\right) = \sum_{i=1}^{n}x_i^2(y_i - \overline{y})$

Formulas for $\hat{\beta}_{2,0}$, $\hat{\beta}_{2,1}$ and $\hat{\beta}_{2,2}$

$\hat{\beta}_{2,0} =$ $\quad \overline{y} - \hat{\beta}_{2,1}\overline{x} - \hat{\beta}_{2,2}\overline{x^2}$

$\hat{\beta}_{2,1} =$

$\hat{\beta}_{2,2} =$

In the spreadsheet add a quadratic regression analysis on the corrected data set. Calculate $\hat{\beta}_{2,0}$, $\hat{\beta}_{2,1}$ and $\hat{\beta}_{2,2}$ and *SSE*. From the latter calculate the coefficient of determination and the standard error of estimate. Include scatter plots of the data plus regression parabola as well as the scatter plot of the quadratic residuals. Fill in the following table.

Use Corrected Data Set	Quadratic Regression	Linear Regression
Constant Term in the Model		
Coefficient on x in the Model		
Coefficient on x^2 in the Model		
Coefficient of Determination		
Standard Error of Estimate		

Fill in the following Regression ANOVA table as shown below.

Source	Sum of Squares	Degrees of Freedom	Mean Square
Quadratic Model			
Error			
Total			

Observed F score _____

At a 1% level of significance, what conclusion do you draw from this ANOVA Table?

Comment on how well the quadratic model fits this data. In particular, how does it compare to the linear model?

What does the pattern of residuals indicate about the adequacy of the quadratic model?

What would be next appropriate step in constructing an adequate model of this system?

Bonus (9 points):

Construct 99% confidence intervals for the parameters of the quadratic regression.

99% Confidence Interval	Lower Limit	Upper Limit
$\beta_{2,0}$		
$\beta_{2,1}$		
$\beta_{2,2}$		

Project 8 **Two Factor Analysis** /48

Name _____ **Due 12/19/2014**

This problem is based on the first major project I did as an engineer at Ray-O-Vac.

Lithium (Li) thionyl choride ($SOCl_2$) batteries offer a great deal of advantages due to their superior energy density (voltage/mass). A problem, however, was the formation of LiCl at the lithium anode, which while it prevents the non-electrical consumption of the lithium, also introduced a "voltage delay" when the battery was discharged across a load. By accident it was discovered that a polymer of cyanoacrylate ("super glue") help in mitigating this effect. A second treatment was to subject each battery after assembly to voltage under load (VUL) test by allowing it to discharge across a 50 ohm resistor for 10 seconds. This procedure was rather cumbersome so it would be difficult to implement in a full scale manufacturing setting. In addition, the amount and placement of the cyanoacrylate within the battery for optimal performance needed to be investigated. In this regard, 16 different pilot assembly runs were conducted. There were 8 different designs of cyanoacrylate amount and application and each assembly was replicated. The different designs were randomly assigned to each pilot assembly. For each batch of batteries produced, half received VUL and half did not. Two response variables of interest were the open circuit voltage (OCV) measured with a potentiometer and the battery capacity. The latter was the total charge delivered in amp hours when the battery was discharged across a 100 ohm load until a 2.7 end point voltage is attained. The results for both response variables are given in the following tables.

OCV	Replication 1		Replication 2	
Design	VUL	No VUL	VUL	No VUL
1	3.623	3.626	3.626	3.632
	3.623	3.632	3.626	3.626
	3.623	3.626	3.626	3.626
2	3.618	3.619	3.619	3.622
	3.615	3.619	3.622	3.626
	3.618	3.619	3.622	3.626
3	3.622	3.618	3.622	3.626
	3.618	3.618	3.622	3.626
	3.619	3.618	3.622	3.626
4	3.622	3.634	3.626	3.634
	3.622	3.632	3.626	3.634
	3.626	3.632	3.626	3.632
5	3.618	3.618	3.622	3.622
	3.618	3.618	3.618	3.622
	3.618	3.622	3.622	3.626
6	3.622	3.622	3.626	3.626
	3.618	3.622	3.622	3.626
	3.618	3.622	3.626	3.626
7	3.622	3.622	3.622	3.622
	3.622	3.622	3.618	3.626
	3.622	3.622	3.618	3.622
8	3.622	3.622	3.622	3.626
	3.622	3.623	3.622	3.619
	3.622	3.626	3.618	3.619

Capacity Factor A	Replication 1		Replication 2	
	VUL	No VUL	VUL	No VUL
1	1.359	1.450	1.283	1.313
	1.473	1.489	1.313	1.271
	1.492	1.433	1.313	1.29
2	1.448	1.490	1.173	1.52
	1.506	1.493	1.511	1.506
	1.478	1.444	1.496	1.505
3	1.472	1.506	1.513	1.502
	1.468	1.519	1.509	1.525
	1.476	1.502	1.531	1.505
4	1.444	1.474	1.342	1.279
	1.464	1.480	1.291	1.268
	1.432	1.464	1.323	1.397
5	1.476	1.504	1.497	1.502
	1.478	1.480	1.508	1.527
	1.488	1.502	1.518	1.524
6	1.433	1.452	1.43	1.468
	1.349	1.392	1.465	1.408
	1.328	1.352	1.419	1.455
7	1.382	1.362	1.426	1.439
	1.357	1.418	1.420	1.337
	1.419	1.511	1.472	1.413
8	1.312	1.334	1.403	1.425
	1.288	1.287	1.442	1.400
	1.311	1.294	1.344	1.453

Perform an analysis which answers the following questions.

1. Is there a significant effect on OCV (alpha = 0.05) due to the design variations?

2. Is there a significant effect on OCV (alpha = 0.05) due to VUL?

3. For OCV is there evidence (alpha = 0.05) of an interaction between design variations and VUL?

4. For OCV is there a significant effect (alpha = 0.05) due to replication?

5. Calculate a 95% confidence interval for the main effect of VUL on OCV.

6. Is there a significant effect on capacity (alpha = 0.05) due to the design variations?

7. Is there a significant effect on capacity (alpha = 0.05) due to VUL?

8. For capacity is there evidence (alpha = 0.05) of an interaction between design variations and VUL?

9. For capacity is there a significant effect (alpha = 0.05) due to replication?

10. Calculate a 95% confidence interval for the main effect of VUL on capacity.

11. Explain why the confidence interval calculation for VUL does NOT use the mean square associated with the VUL factor.