This is a final exam so you can consult your notes and utilize books or whatever online resources you wish, However, please do not consult or collaborate with other classmates. This is meant to test your own knowledge.

1. Name five R's of EDA.

2. In a sample of 10 data, use order statistics to list the five-number summary.

3. Name four goals that may be attained by re-expressing data.

   (a)

   (b)

   (c)

   (d)

4. How do you detect long-tailness? Briefly describe two methods you learned in this course that can help you transform the data.

5. You fit g-h distributions to three datasets, and obtain estimates of the pair (g, h) to be (-0.5, 0.3), (0.5, 0.3) and (1,0.6) respectively. Based on these estimates, how do you compare these data's distribution shapes?

6. A region has been subdivided into 161 population districts with the following population counts. (For your convenience, they have been sorted in increasing order).

```
a =c(1092,1137,1197,1237,1301,1523,1577,
1619,1626,1644,1672,1748,1768,1780,
1796,1816,1843,1844,1902,1919,1983,
1993,2025,2028,2032,2036,2072,2078,
2090,2137,2162,2163,2180,2185,2194,
```

```
2225,2230,2233,2234,2235,2265,2270,
2274,2281,2289,2319,2322,2357,2381,
2398,2421,2421,2443,2522,2549,2552,
2581,2618,2618,2620,2624,2642,2647,
2666,2705,2721,2740,2804,2819,2823,
2860,2873,2906,2913,2926,2929,2931,
2931,2934,2939,2961,3020,3023,3044,
3047,3048,3096,3174,3190,3199,3204,
3222,3225,3278,3287,3292,3300,3339,
3361,3412,3462,3503,3530,3589,3672,
3734,3749,3783,3854,3901,3932,3995,
4001,4006,4118,4134,4320,4346,4385,
4401,4522,4565,4581,4593,4629,4855,
4868,4878,4885,4907,4962,4975,5021,
5127,5155,5160,5183,5229,5242,5379,
5383,5513,5555,5619,5755,5774,5890,
5899,5988,6161,6185,6818,7406,7419,
8175,8220,8282,8827,9027,9042,9805)
```

(a) Construct a qq-plot. Does the distribution appear skewed?

(b) Using the G distribution, estimate A, B, g for these data.

(c) Using the bootstrap with $B = 1000$ simulations, calculate approximate 90% confidence intervals for A, B, g. Also estimate the correlation between the estimates, and construct a pairs plot of the 3 estimates.

(d) Now, using your estimates of A, B, g, transform back to Z and construct a qq-plot of the back-transformed data. How Gaussian do the transformed data look?

(e) Test for normality of the back transformed data using Pearson's goodness of fit test, the correlation of the QQ data test, the Shapiro-Wilk's test and the three ECDF based test statistics (Kolmogorov-Smirnov, Anderson-Darling and Cramer-von-Mises). Discuss the advantages and disadvantages of each of these test statistics.

7. For the following data (n = 161), first construct a QQ plot. Do the data appear to be long-tailed compared to the Gaussian? Now using the theory for the H distribution, fit A, B, h to the data. Construct 90% CIs for each of the parameters using $B = 1000$ bootstrap replications. Also construct the pairs plot for the estimates.

```
b =c(12.87,15.09,17.39,18.62,20.24,23.76,24.35,
24.74,24.81,24.96,25.19,25.75,25.89,25.97,
26.07,26.19,26.35,26.36,26.67,26.76,27.07,
27.12,27.26,27.28,27.30,27.31,27.46,27.49,
27.54,27.72,27.81,27.82,27.88,27.90,27.93,
28.03,28.05,28.06,28.07,28.07,28.17,28.19,
28.20,28.22,28.25,28.34,28.35,28.46,28.53,
```

```
28.58,28.64,28.65,28.70,28.92,28.99,29.00,
29.07,29.16,29.16,29.17,29.18,29.22,29.23,
29.28,29.37,29.40,29.45,29.59,29.62,29.63,
29.71,29.74,29.81,29.82,29.85,29.86,29.86,
29.86,29.87,29.88,29.92,30.04,30.05,30.09,
30.09,30.10,30.19,30.34,30.37,30.38,30.39,
30.43,30.43,30.53,30.55,30.55,30.57,30.64,
30.68,30.77,30.86,30.93,30.98,31.08,31.22,
31.32,31.35,31.41,31.52,31.60,31.65,31.76,
31.76,31.77,31.96,31.98,32.28,32.33,32.39,
32.42,32.61,32.68,32.71,32.73,32.79,33.15,
33.18,33.19,33.20,33.24,33.33,33.35,33.43,
33.60,33.65,33.66,33.70,33.77,33.80,34.03,
34.03,34.26,34.33,34.44,34.68,34.71,34.91,
34.93,35.09,35.40,35.44,36.63,37.81,37.84,
39.47,39.58,39.72,41.00,41.49,41.52,43.50)
```

8. For the data set above, back transform the data and then test for normality using Pearson's goodness of fit test. Use Veleman's rule or Dixon and Kronmal's rule for the number of bins.

9. Generate 100 normally distributed random varaibles with $\mu = 3$ and $\sigma = 2$ using the $d = \text{rnorm}(100, 3, 2)$ command in R. The true mode of the data is 3. Now forget you know that and assume that someone hands you this data. Estimate the mode of the data by finding x value corresponding to the maximum in the Gaussian kernel desity estimator. Now use both bootstrapping and jackknifing to compute the standard error of your estimator for the population mode.

10. Describe how to fit an Robust-Resistant line to data x and y. What is the advantage and disadvantage of fitting an RR line compared to an Least-squares regression line?

11. Briefly explain the difference between Jackknife and Bootstrap methods. Compare them in generating samples, obtaining estimates and the calculation of standard errors.

12. Below are the numbers of great scientific inventions for each year, 1860 to 1959, from the World Almanac (1975):

```
5  3  0  2  0  3  2  3  6  1  2  1  2  1  3  3  3  5  2  4 (=1879)
4  0  2  3  7 12  3 10  9  2  3  7  7  2  3  3  6  2  4  3 (=1899)
5  2  2  4  0  4  2  5  2  3  3  6  5  8  3  6  6  0  5  2 (=1919)
2  2  6  3  4  4  2  2  4  7  5  3  3  0  2  2  2  1  3  4 (=1939)
2  2  1  1  1  2  1  4  4  3  2  1  4  1  1  1  0  0  2  0 (=1959)
```

(a) Tabulate the data (i.e., count the number of years with 0 inventions; with 1 invention; ... with 12 inventions). Based upon a plot of the tabulated data, what distribution might be appropriate for these count data?

(b) Construct a poisson-ness plot for the data along with a plot of the Freeman-Tukey residuals. Estimate from the plot any parameters that are relevant for the distribution. What does the Freeman-Tukey residual plot suggest?

(c) Describe how to calculate the Freeman-Tukey residuals. What do they tell you? What values are "reasonable"?

13. The following data give the carbon dioxide of six plants from Canada and six plants from the southern U.S., measured at 7 different levels of ambient carbon dioxide (CO2) concentration. Each row represents a different level of ambient CO2: 95, 175, 250, 350, 500, 675, 1000. Three plants from each region were chilled overnight and the other three plants were left unchilled. The columns represent the 12 plant/condition combinations: the first two digits indicate the plant number, the 3rd digit indicates the region (1=Canada, 2=U.S.), and the last digit indicates chill (1=unchilled, 2=chilled).

	0111	0211	0311	0412	0512	0612	0721	0821	0921	1022	1122	1222
95	16.0	13.6	16.2	14.2	9.3	15.1	10.6	12.0	11.3	10.5	7.7	10.6
175	30.4	27.3	32.4	24.1	27.3	21.0	19.2	22.0	19.4	14.9	11.4	18.0
250	34.8	37.1	40.3	30.3	35.0	38.1	26.2	30.6	25.8	18.1	12.3	17.9
350	37.2	41.8	42.1	34.6	38.8	34.0	30.0	31.8	27.9	18.9	13.0	17.9
500	35.3	40.6	42.9	32.5	38.6	38.9	30.9	32.4	28.5	19.5	12.5	17.9
675	39.2	41.4	43.9	35.4	37.5	39.6	32.4	31.1	28.1	22.2	13.7	18.9
1000	39.7	44.3	45.5	38.7	42.4	41.4	35.5	31.5	27.8	21.9	14.4	19.9

(a) Median polish this table and data and show the results. Describe the results of the analysis: effects, residuals, patterns, etc.

(b) Approximately how much variability in the data has been explained by the median polish fit? (Calculate a measure and explain how you computed this measure.)

(c) What is the "diagnostic plot"? Explain the calculation of the plot (x-axis, y-axis). Then construct the diagnostic plot. What does the diagnostic plot indicate?

(d) If re-expression is needed, apply it and repeat the analysis: conduct another median polish and calculate a measure of approximately how much variability has been explained this time.

(e) Construct a stem-and-leaf plot of the residuals (use two lines per stem). Do you observe outliers?

(f) Produce box plots to compare residuals among rows and among columns.

(g) Construct the forget-it plot. Describe your conclusions: (i) Which factor is more influential? (ii) Which CO2 level/plant combination has the largest effect?

(h) For your final analysis, let us compute the bootstrap standard errors of the row, column and overall effects from the median polish. To construct the bootstrap you will need to create several bootstrap replicate tables. When you construct each bootstrap replicate table, vectorize and then re-sample with replacement the residuals then recreate a table of residuals. To produce the bootstrap replicate table, then add the row, column and overall effects to bootstrap replicated residual table to produce the bootstrap replicate table. You can perform a median polish on each of the bootstrap replicated tables to obtain the bootstrap standard errors of the median polish coefficients.