# **FSU CIS 5105 (Spring 2021)Systems Performance Measurements and Evaluations**

# **Homework 1 (Due 2/4/2020)**

### Instructor

 Andy Wang (awang@cs.fsu.edu)

 Office: 269 Love Building

 Office Hours: After class and by appointment

This assignment is worth 10% of your course grade. On the due date, please turn in your homework and your Excel file via Canvas. Please name your file in the following format: <last name>-hw1-<last 4 digits of your ID> (5pts).

You are required to show your work for all calculations, and to list the row and column used for all table lookups. It should be possible to duplicate your efforts using a 4-function calculator with a square-root key. You are allowed to use the sum and stdev functions when there are more than ten data points. You may use automated tools for plotting.

1. Eight different benchmarks on a Unix system took 153.4, 255.8, 164.4, 75.1, 206.5, 61.5, 58.3, and 71.1 seconds, respectively.

A. What index of central tendency would you choose, and why? (2pts)

B. Which index of dispersion would you use, and why? (2pts)

C. Calculate the sample mean, median, and mode. (3pts) Please show your work. (3pts)

D. Calculate the geometric and harmonic means. (2pts) Please show your work. (2pts) Are the geometric and harmonic means meaningful? Why, or why not? (2pts)

E. Calculate the range, variance, CoV, and SIQR. (4pts) Please show your work. (4pts)

2. For the same eight benchmarks, the "user" times charged by Unix were 82.8, 0.6, 0.6, 2.4, 30.3, 36.1, 20.2, and 0.4 seconds, respectively.

A. What index of central tendency and index of dispersion would you use, and why? (4pts)

B. Calculate your chosen indices. (2pts) Please show your work (2pts)

3. Triangular distributions are often used as a rough model of an unknown distribution. A triangular distribution has the PDF:

$$f\left(x\right)=\left\{\begin{matrix}0&x<a\\\frac{2}{b-a}\frac{x-a}{c-a}&a\leq x<c\\\frac{2}{b-a}\frac{b-x}{b-c }&c\leq x\leq b\\0&x>b\end{matrix}\right.$$



1. Why is the height of the triangular distribution 2/(b-a)? (1pt)
2. Derive F(x), that is, the CDF of the triangular distribution. (2pts) Please show your work. (2pts)
3. Derive F-1(y) by inverting the CDF of the triangular distribution. (2pts) Please show your work (2pts)
4. Using the inverted CDF derived above, draw a quantile-quantile plot against a triangular distribution with a = 0, b = 3, c = 1, for the following 20 samples (1pt): 3.20, 0.91, 1.10, 4.81, 1.78, 4.38, 2.11, 4.91, 1.93, 3.84, 4.50, 0.96, 0.72, 5.19, 2.61, 4.97, 2.35, 2.06, 4.32, and 2.37. Please describe how you obtained your triangular quantiles (3pts). Does the data appear to match the distribution? (1pt)

4. Seven separate runs of the Modified Andrew Benchmark with NFS-mounted filesystems produced the following real, user, and system times:

 Run Real User System

 1 193.0 82.6 40.4

 2 193.5 82.1 40.9

 3 191.8 83.0 40.2

 4 195.7 82.7 40.5

 5 191.1 82.8 40.3

 6 192.1 82.4 40.4

7 187.1 82.6 40.8

A. Compute the sample mean and standard deviation for each column. (6pts)

B. Calculate the 90% confidence interval for the "system" column (2pts). Please show α (1pt), p (1pt), the table row and column (2pts), and the table value (1pt).

C. Calculate the 99% confidence interval for the "user" column (2pts). Please show α (1pt), p (1pt), the table row and column (2pts), and the table value (1pt).

D. What sample size would be needed to estimate the "real" column to within plus or minus one second at 95% confidence? (1pt) Please show r (1pt), α (1pt), p (1pt), the table row and column (2pts), and table value (1pt).

5. Seven runs of the same benchmark on an experimental distributed file system with four replicas (configuration A) took 175.7, 174.8, 174.9, 174.4, 174.4, 181.9, and 178.4 real-time seconds, respectively.

A. Plot a normal quantile-quantile plot for these values. (1pt) Please describe how you obtained your normal quantiles (3pts). Are they normally distributed? (1pt)

B. Compare these values with the corresponding values given in question 4. Is the performance significantly different from NFS at a 99% confidence level? (1pt) What about at 90% confidence? (1pt) For each confidence interval, please show α (2pts), p (2pts), the table row and column (4pts), and the table value (2pts).

6. Two runs of the benchmark on the same experimental distributed file system with five replicas (configuration B) took 176.8 and 181.4 seconds, respectively. Is the performance for configuration B significantly different from that of configuration A at a 90% confidence level? (1pt) For each configuration, please show α (2pts), p (2pts), the table row and column (4pts), the table value (2pts), and the confidence interval (2pts).

7. The following provides the run times for the benchmark on the same experimental distributed system with one, two, three, and four replicas (the four-replica times are the same as above):

 Run 1-repl 2-repl 3-repl 4-repl

 1 166.8 170.4 184.1 175.7

 2 164.0 170.8 172.5 174.6

 3 167.5 169.8 170.6 174.9

4 166.7 171.0 170.6 174.4

 5 161.3 171.6 171.5 174.4

 6 173.0 168.0 171.0 181.9

7 163.2 165.0 164.1 178.8

A. Fit a linear regression model of run times as a function of replica to this data. (Note: Use the raw data for this fit. Do not calculate the mean of each column and then fit the regression to these means.) (2pts) Please show mean(x) (1pt), mean(y) (1pt), Σ(xy) (1pt), Σ (x2) (1pt).

B. Calculate SSY, SS0, SSE, SST, and SSR. (5pts)

C. Calculate R-squared and the correlation R(x,y). (2pts)

D. Calculate the standard deviation of the errors. (1pt)

E. Calculate the confidence intervals for the regression parameters b0 and b1 at the 95% level. Please show α (1pt), p (1pt), the table row and column (2pts), the table value (1pt), and confidence intervals (4pts). Are the parameters significant? (2pts)

F. For each of the four replication levels, plot the mean of the seven observations on a scatter plot, with the number of replicas as the X coordinate. (1pt) About each point on the graph, plot a vertical line showing the 95% confidence interval for that point. (1pt) On the same graph, plot the fitted line (1pt) and two dashed lines (2pts) representing the confidence intervals for the fitted parameters. The final graph should show the means of the raw data, their confidence intervals, the fitted line, and the confidence intervals for the fitted line, all on the same plot.

G. Using the regression model, predict a mean and 95% confidence interval for a future sample of two observations at the five-replica level. Please show m (1pt), xp (1pt), se (1pt), the standard deviation for mean of the future sample m (1pt), α (1pt), p (1pt), the table row and column (2pts), the table value (1pt), and confidence interval. (2pts) How does this compare with the mean of the two actual observations from question 6? (1pt)

H. Based on the scatter plot from part (F), does the linear model appear to be a good fit? (1pt)

I. Calculate the errors in the regression and prepare a scatter plot of the errors (y-axis) against the predicted response. (1pt) Is there a visible trend to the errors? (1pt) Does the data appear to be homoscedastic? (1pt)

J. Prepare a scatter plot of the errors (y-axis) against the run number. (1pt) Is there a trend? (1pt)

K. Prepare a normal quantile-quantile plot of the error residuals. Please describe how you obtained your normal quantiles (3pts). Are the errors normally distributed? (1pt)

1. Please submit a recent photo (e.g., a Xerox copy of your face) and tell me a favorable quote and why it is memorable to you. (5pts)