

Stat 6601 Midterm

- Each problem is 10 points unless stated otherwise
- Make MS Word document with name "last_name,first_name.doc" (e.g. "kwon, jaimie.doc") from the template available on the course website.
- Include a) typed answers, b) cut-and-pasted R codes and R-outputs, and c) R-plots.
- Attach the file to an email with subject line "stat 6601, midterm" (nothing else please; no message body)
- Email your answers to me no later than Midnight, Tuesday November 2nd. Results submitted later than then won't be graded.
- Minimize comment in the MS Word document. Typical headings (title, date, and author) and question numbers are enough. (Please keep in mind that I have to go over 50 of them!)

#1. (110 pt + 20 pt) Use the air conditioning data again. (See the lecture node):

```
y <- c(3, 5, 7, 18, 43, 85, 91, 98, 100, 130, 230, 487)
n <- length(y)
```

We are interested in estimating $\theta = \log(\mu)$, the log mean inter-failure time.

- 1) Compute $T = \log(\bar{Y})$ for the sample. What's the value t ?
- 2) Write R codes that simulate a single sample y^* from the fitted parametric model $Exp(1/\bar{Y})$ and compute $t^* = \log(\bar{y}^*)$. What's the value of t^* ?
- 3) Write R codes that simulate $R=1000$ samples y^* from the fitted parametric model $Exp(1/\bar{Y})$ and compute $t^* = \log(\bar{y}^*)$. Store the values of t^* in "t.star".
- 4) Draw probability histogram of the t^* simulated in 3 with 50 bins. On top of the probability histogram, draw the kernel density function estimate. Using the default bandwidth is OK, though you are free to specify bandwidth that makes the curve look more reasonable.
- 5) Draw normal Q-Q plot of the t^* . Also draw a reference straight line using "qqline" function. What can you say about the distribution of t^* and T from the visual inspection of the normal Q-Q plot? Is it close to normal?
- 6) Independent of the inspection in 5), one decided to use normal approximation to the distribution of $(T - \theta) = \log(\bar{Y}) - \log \mu$. In that case, one could estimate the bias and variance of the statistic T using the bootstrap simulation results obtained in 3). What are the values of the estimated bias and variance? Assign them to 'B' and 'V' respectively in R.
- 7) Using the estimate bias and variance (B and V) obtained in 6), combined with the normal approximation, compute the 95% confidence interval for $\theta = \log \mu$.
- 8) Now use nonparametric bootstrap to answer above questions. Write R codes that simulate $R=1000$ samples y^* from the empirical distribution of y (Use 'sample' function in R.) and compute t^* . Store the values of t^* in "t.star.np".
- 9) Repeat 4) above (histogram + kernel density estimate) for the nonparametric bootstrap simulation obtained in 8). Also draw the normal Q-Q plot. Does the distribution look normal?

- 10) Compute the basic bootstrap confidence interval with confidence coefficient 95% using the simulation obtained in 8). Note that you can use 'quantile' function to compute $t_{((R+1)\alpha)}^*$ etc.
- 11) Draw the histogram of 9) again. On it, use "abline(v=...)" function in R to mark boundaries of the 95% confidence interval from the normal approximation obtained in 7) in red vertical lines. Also mark 95% basic bootstrap confidence interval obtained in 10) in blue vertical lines. Which one is wider?
- 12) (Extra credit) Write R code to compute the number of t^* obtained from the nonparametric bootstrap 8) that fall in the 95% confidence interval of 7). Divide it by R to obtain the proportion. This is a good estimate of the coverage probability of the normal approximation. What's the value?
- 13) (Extra credit) If \hat{G}_R is a good approximation of the distribution of G , the distribution of T , what does the result in 12) say about the advantage of nonparametric bootstrap? Note that by definition, nonparametric bootstrap confidence intervals obtained as in 10) ALWAYS have correct coverage probabilities.

#2. (50 pt + 20 pt) Suppose $n=100$ independent observations of $Y_j = (X_j, Z_j)$ are observed. It's stored in "corr.dat" on the class webpage.

- 1) Use "read.table" function in R to store it in 100 by 2 matrix "data". Use "cor(data[,1], data[,2])" or "cor(data\$x, data\$y)" function to compute the sample correlation coefficient.
- 2) One can perform a single nonparametric simulation by the code "data.star <- data[sample(1:n, replace=TRUE),]". Use this to write R codes that simulate a single sample y^* and compute the correlation coefficient ρ^* .
- 3) Write R codes that simulate $R=1,000$ samples y^* and compute correlation coefficient ρ^* for each of them. Store the results in "rho.stars".
- 4) Draw histogram + kernel density estimate (overlaid) of ρ^* .
- 5) Compute the basic bootstrap confidence interval with confidence coefficient 95% using the simulation obtained in 4). Note that you can use the 'quantile' function.
- 6) (Extra credit; 20 pt) Perform parametric bootstrap to obtain the same basic bootstrap confidence interval with confidence coefficient 95%. Use "mvrnorm(n, c(mu1, mu2), matrix(c(sigma11, sigma12, sigma21, sigma22),2,2))" function in MASS library.