project/term paper:

Deadline for submission November 29

[A complete report and R codes with appropriate comments in separate files]

Towards this assignment, each of you have been assigned a (separate) distribution denoted by P. Given below are some guidelines/questions that should be addressed in the term report. However, the idea of the exercise is to push you in the direction of research and hence you are advised to use the freedom liberally in that direction, keeping the guideline in mind. It is possible that we would schedule a 15 min presentation in which you would be required to present ONLY the more interesting results.

Part 1: Study of the population characteristics

1. For general set of parameters (θ, could be vector) and obtain expressions /compute different measures of central tendency, dispersion, skewness etc.

The list of characteristics (γ) should include (but not be limited to) mean, median, another percentile (say 95^{th}) inter-quartile range, standard deviation, some proportion (e.g. proportion of values > mean +s.d.)

(These would be readily available in many texts/internet sources – you may derive as many of these analytically as possible/ as you like, or compute numerically, in special cases).

Among other things, this part of the study should investigate,

- * How do these characteristics (γ) of interest change with θ?
- How often would a Box-plot detect outliers (major & minor) if the actual distribution is P?
- * Can you think of any "nonconventional" measures of central tendency and dispersion? Comment on their appropriateness with the specific distribution P.
- Consider X_1, X_2, X_3 drawn from your population. What is the probability of the smallest of the observations exceeding population median? What happens if sample size is increased from $n=3$?

Part 2: Data

2. Write a small code in R, to simulate data from this population (with specific choice of parameters, which can be altered) --- You would need to use this code in later part, at times repeatedly. (In most cases, there are inbuilt routines – so you would just need to look for the correct function --- you should not spend much time on this anyway.) We will refer to this as "simulated data".

3. Get hold of a "real data" set for which P will pass any of the goodness of fit tests --- thus providing basis for using P as a reasonable model for this data.

4. For the study of robustness study we will consider a contaminated /mixture distribution --- P contaminated with some other distribution. Prepare a code for generating data from this.

Part 3: Sampling distribution

5. Consider reasonable estimators T's corresponding to the different γ's. (**OF PARTICULAR INTEREST IS VARIANCE AND/OR OR STANDARD DEVIATION**) We would want to study the sampling distribution of T's. (Note that these are different problems corresponding to the different γ's. So start with one, and then repeat the exercise for the other characteristics). Among other things, the goal would be to get an idea of

   - Distribution for some selected small sample(s) (e.g. n=5)
   - Is the asymptotic distribution normal? Explore around what sample size does the asymptotic kick in?
   - * Can you discover a "good" known structure for the distribution for some small/moderate sample sizes?
   - Compare the (actual) theoretical distribution with simulated distribution, whenever possible. For example, you should be able to find out the actual distribution of order statistics (using P) – compare this with its simulated distribution. What is the role of simulation size? [Simulation size = No. of samples of given sample size --- it should be "very large".]

     For these parts you would require to expand/repeatedly use code generated in 2.

   - Study the exact and/or asymptotic sampling distribution of sample IQR (and another statistic of your choice based on order statistics, other than median), if the sampling is done from your population.

Part 4: Statistical Inferences

6. Find the MLE of θ. Find the MLE's of the other characteristics of interest, γ . What can you say about the (asymptotic) distributions of these estimators? In particular, focus on the asymptotic variances? [You would need to calculate Fisher's information]

7. For at least some of the γ's, can you think of more than 1 competing "good" estimators? Compare them on the basis of

   - Bias
   - Variance

- Mean square error

8. A) Find the confidence intervals for the γ's, in case the sample sizes are suitably large. B) For one of these C.I.'s of a suitable γ, comment / study the accuracy of the confidence coefficient with increase in sample size. C) * For one of these γ's, obtain (at least approximately accurate) C.I. for some "small" sample sizes as well? (Hint: Use simulation to obtain the distribution of the relevant statistic in the first stage)

9. Formulate simple null vs simple alternative hypothesis involving γ's, where γ= a) mean b) variance c) proportion of some kind (say P(X>t) for some t d) some percentile of the distribution. Find the MP tests using N-P lemma, wherever possible. Obtain very specific forms (values), given specific numeric choices of $\gamma 0$ and $\gamma 1$.

    - For (at least) two of these parameters, obtain the explicit expression of the critical region as well as the power of the MP test. Note that, for this you would need the null distribution (distribution of the test statistic under null hypothesis) as well as the distribution of the TS under alternate hypothesis. If these distributions are theoretically not valid (explore if asymptotic distributions are valid), then even for small/moderate sample sizes, you should be able to derive the distribution by simulation. Ideally one of the computations could involve that.

    - Check on (at least) couple of cases, if the UMP tests exist for composite alternative hypothesis

10. In this part, we wish to compare efficiency and robustness of two estimators, one parametric (MLE) and the non-parametric (empirical/sample percentiles). Let γ be some chosen fractile/percentile of the population parameter, i.e. ξ_p, like median/ third-quartile/ or VaR. (We would ideally want to change p, and study the impact). [You may have addressed some initial parts of this in earlier steps like 7etc)

    - In class, we have seen theoretical asymptotic distributions for both estimators. Calculate theoretical values of asymptotic variances of the two estimators. Using simulation data from the population, explore (a) when the asymptotics kick in both in terms of distributional results as well as validity for approximate variance (standard errors) of the two estimator. In summary, compare the efficiency of the NP estimate, in comparison to the MLE.

    - In this part, we wish to investigate, how much the NP estimator out performs the MLE, if the population is mis-specified. Towards this, you would consider a mixture model, the population being $F_\varepsilon = 1-\varepsilon$ * F + ε * G, with F being the model distribution you are considering and G being contamination distribution (sufficiently different from F), and $0 < \varepsilon < 1$ being contamination percentage. MLE would be based on mis-specified knowledge of F only, and of course the NP estimative does use the population structure

at all. From simulation data generated from F_ε, how/when NP estimate outperforms MLE in terms of bias/variance/MSE as a function of n, p, ε.

Carry out one or more of the following

11. This is to find out which among the ANOVA (F-Test) or Kruskal-Wallis test is more appropriate for your population. Towards this, consider testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, for some k between 3 to 5, and consider three situations (a) all sample sizes between 10 to 20 (b) all sample sizes between 30 and 60 (c) all sample sizes between 100 to 250. Consider the other (nuisance) population parameters to be identical or such that the population variances are equal (but unknown). Consider a few (say 3-4) specific choices for the alternative hypothesis. Take a specific choice of size (alpha), compare the power of the two test procedures. Do this analytically or through simulation or by combining the two approach, as you find appropriate.