ST204 Nonparametric Statistics
2021-22 Semester 2
Assignment Sheet 5

Due at 16:00 on Friday 6th May 2022.

Only one randomly chosen question will be marked. Submit answers to questions not already covered in tutorials only. Your submission file should be in the form of a merged .pdf and your code must be provided for any questions involving R. Otherwise, you are free to mix typed/handwritten solutions as you see fit.

1. A random sample of test scores (%) for an exam were recorded.
   The score values are: 29,41,52,56,61,62,64,67,72,75,78,79. Use R to answer the following:
   (a) Describe the distribution, providing an appropriate graph and summary statistics as support.
   (b) Use the sign test approach to construct a confidence interval for the population median θ. Explain why the sign test is more appropriate than the Wilcoxon signed-rank test in this case.
   (c) Use a bootstrapping approach to construct a CI for the population median θ by explicitly writing out a `for` loop in R without using any additional packages. List the steps involved in doing so (in the context of this example and in terms of the underlying process, not only in terms of how it was programmed in R). Provide a histogram of the bootstrap medians in your answer with vertical lines to indicate the sample median and the bootstrap CI bounds.

2. It is believed that some psychological therapies can help teenage girls suffering from anorexia to *gain* weight. In a study, 29 teenage anorexic girls received cognitive behavioural therapy. The weight change over the course of the treatment (after minus before) was recorded for each girl as follows:

6.7	0.7	-0.1	-0.7	-3.5	14.9	3.5	17.1	-7.6	1.6	11.7	8.8
6.1	1.1	-4.0	2.9	-9.3	2.1	1.4	-0.3	-3.7	-1.4	-0.8	-6.6
2.4	12.6	1.9	3.9	0.1	8.4	-0.7	6.3	5.8	9.9	-11.1	7.1

   (a) Between the sign test and Wilcoxon signed-rank test, which is more appropriate for testing hypotheses about the population median for these data? Explain why and state clearly the null and alternative hypotheses.
   (b) Use a bootstrap approach (with 3000 bootstrap samples) and the method identified in part (a) to test these hypotheses. Include a histogram of the bootstrap values with the true sample value indicated by a vertical line and comment on how the graph relates to your p-value.
   (c) Briefly describe how you did the bootstrap sampling and why you did it this way, in the context of this example (in terms of underlying theory, not only how it was programmed in R).
   (d) Give one reason why using a bootstrap approach is reasonable in this case, rather than relying on direct application of the test.

3. The effective channel length (in microns) is measured for 1225 field effect transistors. The data is recorded in `CA5_Transistor.csv`. Use kernel density estimation to find a suitable density estimate for this data.
   (a) First, generate a density histogram for the length data. Use `freq=FALSE` and `breaks=16`.
   (b) Then, examine various bandwidths to identify a suitable value (using the Gaussian kernel). Justify your choice in writing.
   (c) Examine various kernels to identify a suitable one. Ensure that you also try some of the kernel options that were **not** discussed in class. Justify your choice in writing.
   (d) Superimpose the chosen density function on the histogram generated in (a).
   (e) Overlay a parametric density estimate assuming a Gaussian distribution and comment on its limitations for these data.

4. A random sample of size $n = 14$ is reported below:

   4, 7, 14, 12, 13, 8, 19
   7, 5, 6, 9, 21, 11, 16

   Evaluate the density at $x = 9.5$, *by hand*, using a Gaussian (standard normal) kernel with a bandwidth of 2.5. Give your final answer to at least 5 decimal places. You must show all your workings.

5. For the same set of data as Q4, evaluate the density at $x = 9.5$, *by hand*, using an Epanechnikov kernel with a bandwidth of 4.5. Your final answer, in the form of a fraction, should be exactly 38/567. You must show all your workings.