Stat 3005 – Nonparametric Statistics

Final project

Time limit: 72 hours

Fall 2024

Instructions

1. The full mark of the project is 100.
2. There is an optional bonus question that is worth 10 marks. (E.g., if you get $95+10$, your project score is 100.)
3. Complete the project solely by yourself. No collaboration is allowed. You may consult lecture notes, tutorial notes, homework assignments, quizzes, textbooks, online materials, etc, however, you cannot solicit or obtain assistance from or provide assistance to other people in any form for any specific content on the project. Activities considered cheating include, but are not limited to, copying, rephrasing or modifying contents from websites, discussing project questions with other people, and asking for help on Internet forums. All cheating cases will be passed to the the Senate Committee on Student Discipline with no exception.
4. Generative AI tools (e.g., ChatGPT) can be used for auxiliary purposes so that the answers reflect your own intellectual ideas and independent effort. The graders have the ultimate power to decide the appropriateness and the level of originality of the submitted answers.
5. Quote the results that appear in the lecture notes if you wish to use them, e.g., write "by Theorem 3.1", "by the CLT for rank statistics", etc.
6. Prove the results that do not appear in lecture notes if you wish to use them.
7. Define your abbreviations clearly unless they have been defined in the lecture notes.
8. Submit the following two documents to Blackboard before the deadline.
    - Written part. Either (i) handwrite and scan your answers, or (ii) type your answers in LaTeX. Use A4-size papers. Compile your answers in one single `.pdf` file. Sign the Honor Code below, and attach this page as a cover of your submitted file. Name the document in the format `S3005_F_sid_name.pdf`. E.g., `S3005_F_1155001234_ChanKinWai.pdf`. Note that all plots, numerical answers, simulation results, etc must be included in the written part. Graders will not run your submitted codes to check the answers.
    - Computation part. Save your `R` codes in one single `.r` file and name it in the formate `S3005_F_sid_name.r`. E.g., `S3005_F_1155001234_ChanKinWai.r`.

   Submission without properly signed honor code will not be graded. Submissions with incorrect formats will not be graded. Unclear or blurred answers will not be graded. Late submission will not be graded. You may re-submit your answers as many time as you wish, however, only the last submission will be graded.

Honor Code

Please read the Honor Code below and sign your name.

> *The Chinese University of Hong Kong places very high importance on honesty in academic work submitted by students, and adopts a policy of zero tolerance on academic dishonesty. All work presented in this project, including ideas, discoveries, interpretations, codes, derivations, and conclusions, should solely reflect independent effort, and strictly adhere to the standard of academic integrity and the instructions above (see the academic honesty guide provided by CUHK for more details).*

I understand the above instructions, and affirm that all work conforms to the standards of the Honor Code.

Signature: ______________________

Print Name: ______________________

Student ID: ______________________

Date: ______________________

Question 1 (★☆☆ — Hypothesis test (40%)). Let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\theta_x, \theta_x^2)$ and $Y_1, \ldots, Y_n \overset{\text{IID}}{\sim} \mathrm{N}(\theta_y, \theta_y^2)$ be two independent samples, where $\theta_x, \theta_y \in \mathbb{R}$ are unknown, and $n \geq 2$. Ken claimed that θ_x and θ_y are not the same. The dataset `Q1.csv` is used. For each test below:

(i) Prove or disprove that it has a correct size under the null hypothesis.

(ii) Argue whether it has reasonably high power under the alternative hypothesis. Only soft arguments are needed. Explain your answer in $\lesssim 50$ words. Simulation may be used for the explanation.

(iii) Perform the proposed test and state the testing result at size 5% if the proposed test is correct.

Report your results using the following table by placing (✓) in the appropriate cells:

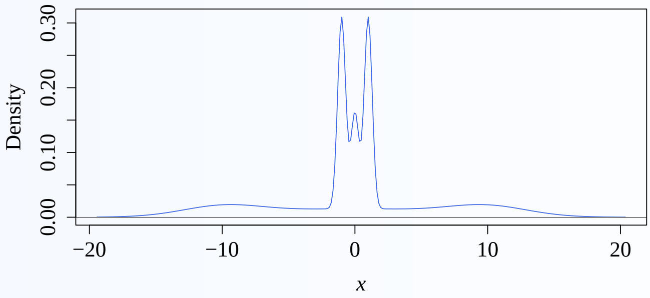
Test	Arnold	Brian	Calvin	Daniel	Edwin	Francis	Gavin	Hannah
(i) has correct size								
(ii) has high power								
(iii) reject the null								

1. (5%) Arnold proposed using the rank sum test.
2. (5%) Brian proposed using Ansari–Bradley test.
3. (5%) Calvin proposed rejecting H_0 if either Arnold or Brian rejected H_0.
4. (5%) Daniel proposed using Cramér–von Mises test.
5. (5%) Edwin proposed using the R code: `ks.test(x, ecdf(y))` to perform the test, where `x` and `y` are vectors storing $X_{1:n}$ and $Y_{1:n}$, respectively.
6. (5%) Francis proposed using a permutation 2-sample t-test based on group permutation.
7. (5%) Gavin arranged the data in pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and used the exact sign-rank test.
8. (5%) Hannah proposed finding a 95% bootstrap confidence interval (CI) for the variance difference $\theta_x^2 - \theta_y^2$. If the CI does not contain 0, then H_0 is rejected.

Question 2 (★★☆ — Adaptive bandwidth (30%)). Let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} f$, where f is a PDF having nice enough properties including absolutely continuity of f'', finiteness of $\int_{-\infty}^{\infty} \{f'''(x)\}^2 \, \mathrm{d}x$, etc. In the lecture note, the kernel density estimator (KDE)

$$\widehat{f}_\ell(x) = \frac{1}{n\ell} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\ell}\right)$$

is equipped with a fixed bandwidth ℓ across x, where K is a kernel. In some applications, $f(x)$ has variable smoothness across x, thus, it may be more appropriate to use a variable bandwidth at different x. An example of such density is shown below. This "stretching pose" density is generated from a mixture of seven normal distributions, each with means $(-10, -3, -1, 0, 1, 3, 10)$, standard deviation (3,6,0.3,0.3,0.3,6,3), and weight $(1, 1, 2, 1, 2, 1, 1)/9$. The dataset `Q2.csv` of size $n = 3005$ is generated from this density. This dataset is used in this question.



You may use the following R code to produce the above density plot:

```r
mu =     c(-10, -3, -1, 0, 1, 3, 10)
sigma = c(3, 6, 0.3, 0.3, 0.3, 6, 3)
p =     c(1, 1, 2, 1, 2, 1, 1)/9
t = seq(from=-20,to=20,length=301)
f = colSums(simplify2array(lapply(t,function(t)dnorm(t, mu,sigma)))*p)
plot(t,f,type="l")
```

Throughout this equation, we denote the rth order derivative of a function $g(x)$ with respect to x by $g^{(r)}(x)$ for $r = 0, 1, 2, \ldots$.

1. (10%) It is given that $\widehat{f}_\ell^{(r)}(x)$ can be used to estimate $f^{(r)}(x)$ for $r = 0, 1, 2, \ldots$. You may use without proof that the asymptotic mean integrated squared error-optimal ℓ for $\widehat{f}_\ell^{(r)}(x)$ is

$$\ell_r^* = \left[\frac{(2r+1)\lambda_K^{(r)}}{n\sigma_K^4 \int_{-\infty}^{\infty}\{f^{(r+2)}(t)\}^2\,\mathrm{d}t}\right]^{\frac{1}{2r+5}}, \quad \text{where} \quad \lambda_K^{(r)} = \int_{-\infty}^{\infty}\{K^{(r)}(t)\}^2\,\mathrm{d}t, \qquad \sigma_K^2 = \int_{-\infty}^{\infty} t^2 K(t)\,\mathrm{d}t.$$

   (a) Prove that

$$\widehat{f}_\ell^{(r)}(x) = \frac{1}{n\ell^{r+1}}\sum_{i=1}^{n} K^{(r)}\left(\frac{x - X_i}{\ell}\right). \tag{2.1}$$

   (b) The rule-of-thumb bandwidth under the normal reference rule is found by assuming $X_i \sim \mathrm{N}(\mu, \sigma^2)$ and K is the standard normal density. Under this rule, verify that $\ell_0^* = 1.059\sigma/n^{1/5}$ and $\ell_2^* = 0.940\sigma/n^{1/9}$. Numerical methods can be used for finding the leading coefficients in ℓ_0^* and ℓ_1^*.

   (c) If the σ in part (b) is unknown, it can be estimated by

$$\widehat{s} = \min\{\widehat{\sigma},\ \widehat{\mathrm{IQR}}/1.34\}.$$

   where $\widehat{\sigma}^2$ is the sample variance and $\widehat{\mathrm{IQR}}$ is the sample interquartile range. Write an R function `kdeDerivative = function(xeval, data, r=0)` for computing $\widehat{f}_\ell^{(r)}(x)$ under the normal reference rule, where

   - `xeval` is a vector containing all values of the function input x in (2.1),
   - `data` is a vector storing the dataset $X_1, \ldots, X_n$, and
   - `r` is an integer indicating the order of the derivative r.

2. (10%) Jones (1990) proposed a method which is now known as a *balloon estimator*. It estimates $f(x)$ as

$$\widehat{f}_{\mathrm{BAL}}(x) = \frac{1}{n\widehat{\ell}(x)}\sum_{i=1}^{n} K\left(\frac{X_i - x}{\widehat{\ell}(x)}\right),$$

   where $\widehat{\ell}(x)$ is an estimator of the optimal bandwidth $\ell^*(x)$ at x given in (8.8) of the lecture note, i.e.

$$\ell^*(x) = \left(\frac{f(x)\lambda_K^{(0)}}{n\{f''(x)\}^2\sigma_K^4}\right)^{1/5},$$

   One may estimate $\widehat{\ell}^*(x)$ by using `kdeDerivative` in part 1, and use it to compute pilot estimators of $f(x)$ and $f''(x)$. Write a R function `kdeBallon = function(xeval, data)` for computing the balloon estimator $\widehat{f}_{\mathrm{BAL}}$. Compare this estimate with the true density and the standard KDE using different types of non-adaptive bandwidths.

3. (10%) Breiman, Meisel, and Purcell (1977) proposed another method called *sample-point estimator* defined as

$$\widehat{f}_{\mathrm{SAM}}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\ell_i}K\left(\frac{X_i - x}{\ell_i}\right),$$

   for some $\ell_1, \ldots, \ell_n$. Later Abramson (1982) suggested using $\ell_i \propto f(X_i)^{-1/2}$. In particular, we consider

$$\ell_i = \ell_0\left(\frac{\bar{f}_{\mathrm{GM}}}{\widehat{f}_{\ell_0}(X_i)}\right)^{1/2}, \qquad \text{where} \qquad \bar{f}_{\mathrm{GM}} = \left(\prod_{i=1}^{n}\widehat{f}_{\ell_0}(X_i)\right)^{1/n}$$

   is the geometric mean of the pilot estimates of $f(x)$ evaluated at the sample points $X_1, \ldots, X_n$ with the plug-in bandwidth ℓ_0 computed by the R function (`bw.SJ`). Write a R function `kdeSample = function(xeval, data)` for computing the sample-point estimator $\widehat{f}_{\mathrm{SAM}}$. Compare this estimate with the true density and the standard KDE using different types of non-adaptive bandwidths.

Question 3 (★★★ — Kruskall–Wallis test (30%))**.** Consider a k-sample location problem, where $k \geq 2$. Let

$$X_{ij} = \mu + \tau_j + \varepsilon_{ij}, \qquad i = 1, \ldots, n_j, \qquad j = 1, \ldots, k,$$

where μ is the overall location, τ_j is the jth "treatment" effect, ε_{ij}'s are IID continuous RVs, and n_j is the size of the j sample. Denote $N = n_1 + \cdots + n_k$. The goal is to test

$$H_0 : \tau_1 = \cdots = \tau_k \qquad \text{against} \qquad H_1 : \exists\, j, j' \text{ such that } \tau_j \neq \tau_{j'}.$$

So, H_0 means that all k samples have the same location. In 1918, R. A. Fisher first proposed a method called analysis of variance (ANOVA) for testing H_0 upon assuming that $\varepsilon_{ij} \overset{\text{IID}}{\sim} \mathrm{N}(0, \sigma^2)$ for some $\sigma > 0$. The test statistic is

$$F = \frac{\frac{1}{k-1}\sum_{j=1}^k n_j(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2}{\frac{1}{N-k}\sum_{j=1}^k \sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\cdot j})^2} = \frac{\frac{1}{k-1}\mathrm{SSB}}{\frac{1}{N-k}\mathrm{SSW}}, \quad \text{where} \quad \bar{X}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} \quad \text{and} \quad \bar{X}_{\cdot\cdot} = \frac{1}{N}\sum_{j=1}^k \sum_{i=1}^{n_j} X_{ij}.$$

So, F can be interpreted as the ratio of sample variance between groups ($\frac{1}{k-1}\mathrm{SSB}$) and the sample variance within groups ($\frac{1}{N-k}\mathrm{SSW}$). It is worthwhile to note that the sum of SSW and SSB satisfies the identity: $\mathrm{SST} \equiv \sum_{j=1}^k \sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = \mathrm{SSB} + \mathrm{SSW}$. Under H_0, $F \sim F_{k-1,N-k}$, i.e., the F-distribution with degrees of freedom $k-1$ and $N-k$. On the other hand, under H_1, the between group variance is expected to be larger than the within group variance, hence F is large. So, we reject H_0 at size α if $F > \texttt{qf}(1-\alpha, k-1, N-k)$. A rank-based version of ANOVA is called Kruskall–Wallis test, which only assumes ε_{ij}'s are IID continuous RVs. Define R_{ij} be the rank of X_{ij} among the entire data set of size N. The Kruskall–Wallis test statistic is defined as

$$T = \frac{12}{N(N+1)}\sum_{j=1}^k n_j(\bar{R}_{\cdot j} - \bar{R}_{\cdot\cdot})^2, \quad \text{where} \quad \bar{R}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{ij} \quad \text{and} \quad \bar{R}_{\cdot\cdot} = \frac{1}{N}\sum_{j=1}^k \sum_{i=1}^{n_j} R_{ij}.$$

1. (5%) Which test statistic is nonparametric, F or T? Explain. Use $\lesssim 50$ words.
2. (5%) Prove that

   $$T = -3(N+1) + \frac{12}{N(N+1)}\sum_{j=1}^k n_j \bar{R}_{\cdot j}^2.$$

   Hence, or otherwise, argue that T is a distribution-free statistic under H_0.
3. (5%) Argue that T is expected to be large under H_1. *(Hints: compare it with F.)*
4. (5%) Let $s(i) = S(\frac{i}{N+1})$ for some non-constant $S : [0,1] \in \mathbb{R}$ such that $\int_0^1 S^2(u)\,\mathrm{d}u < \infty$. Let $\mu_s = \sum_{i=1}^N s(i)/N$ and $\sigma_s^2 = \sum_{i=1}^N \{s(i) - \mu_s\}^2/N$. It is given that if X_{ij}'s are IID continuous RVs, then

   $$Q = \frac{1}{\sigma_s^2}\sum_{j=1}^k \frac{1}{n_j}\left[\sum_{i=1}^{n_j}\{s(R_{ij}) - \mu_s\}\right]^2 \xrightarrow{\text{d}} \chi^2_{k-1} \qquad \text{as} \qquad \min(n_1, \ldots, n_k) \to \infty.$$

   Find the limiting distribution of T under H_0.
5. (5%) Propose a permutation procedure for performing Kruskall–Wallis test. Explain why is it a valid permutation for this problem. Write down the maximum possible number of permutation for this procedure.
6. (5%) Write a R function `kw.test` to perform (i) exact, (ii) asymptotic, and (iii) permutation Kruskall–Wallis test. Compute these three p-values based on the dataset `Q3.csv`.

Question 4 (★★★ — Bonus (10%))**.** You may do whatever you think is relevant and insightful for understanding one and only one of Questions 1–3. The question number (1/2/3) must be stated clearly in the first line of your answer. The answer must be highly relevant to Stat 3005. The main part must be within one page, including all graphs and tables but excluding codes and supplementary materials. The font size and margin size should be sensibly chosen to make the answers readable. High scores are given to insightful and nicely presented answers.