Stat 3005 – Nonparametric Statistics

Mid-term project

Time limit: 48 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_M_sid_name.pdf`. E.g., `S3005_M_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_M_sid_name.r`. E.g., `S3005_M_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 (★☆☆ — Theory of rank-type test (40%)). Let the data $X_1, \dots, X_n$ be generated as follows:

$$X_i = \mu + \sigma_i \varepsilon_i, \qquad i = 1, \dots, n,$$

where $\mu \in \mathbb{R}$ is a constant, and $\varepsilon_1, \dots, \varepsilon_n$ are IID continuous RVs. Denote the rank of X_i by R_i. Assume that n is an odd integer. Kimberley suspected that $\sigma_1 > \sigma_2 > \cdots > \sigma_n$. She proposed the following test statistic:

$$T = \sum_{i=1}^{n} i \left| \frac{R_i}{n+1} - \frac{1}{2} \right|.$$

1. (10%) Formally state the null and alternative hypotheses for testing her claim.
2. (10%) Derive $\mu_T := \mathsf{E}(T)$ and $\sigma_T^2 := \mathsf{Var}(T)$ under your null hypothesis.
3. (10%) Derive the limiting null distribution of T after proper standardization.
4. (10%) Argue that T is a sensible test statistic, and state the testing procedure. (Use $\lesssim 100$ words.)

See Remark 1 for hints.

Question 2 (★★☆ — Sex differences in sports performance (30%)). Suppose you are Anthea, a statistician working in the research department of a sport company. You are responsible for offering statistical consultation services to your colleagues. You received the following email from your colleague, Kathy.

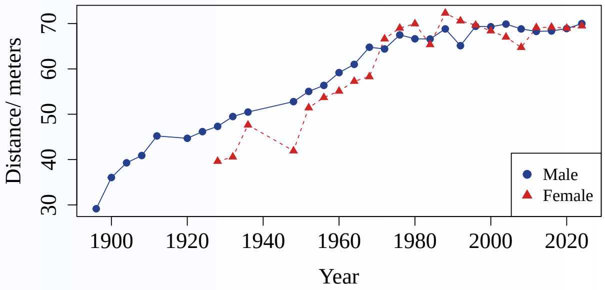
> Dear Anthea,
>
> I am writing to ask for statistical help with my analysis on sex differences in sports performance. In many sports that depend on muscular power, it is commonly believed that male athletes outperform female athletes on average. However, we believe that the advantage may not be as pronounced in some sports. I am studying the Olympic winning distances in the discus throw, but I am unsure how to conduct a proper statistical analysis. Could you kindly assist me in analyzing the data?
>
> We collect the winning distances (in meters) of the gold medalists in the men's and women's discus throw at the Summer Olympics from 1896 to 2024. The dataset was downloaded from Statista and BBC News. I have conducted some preliminary analysis; see the attache file. In the dataset `discus.csv` attached, approximately 60% (14 out of 23) of the winning distances in the men's discus throw are longer than those in the women's discus throw, suggesting that the chances are roughly equal. After inspecting the dataset, I found that starting in 1976, female gold medalists have performed very well. Therefore, using the dataset from 1976 onward, we conducted a t-test and failed to reject the null hypothesis of equal means for men's and women's winning distances. I am not sure if the analysis makes sense or not.
>
> I would greatly appreciate your statistical insight and wisdom. Thank you so much.
>
> Best,
> Kathy

The attachment contains the following information. Kathy's dataset can be downloaded in R as follows.

```
id = "1bhOLQ8QzJu6P_dM7UHYK1-6zFLntaAgV"
data = read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id))
```

The first few rows of the dataset `discus.csv`, the time series plots, and some analyses are shown below:

```
> head(data,8)
               Year   Men Women
1       Athens 1896 29.15    NA
2        Paris 1900 36.04    NA
3   St. Louis 1904 39.28    NA
4       London 1908 40.89    NA
5   Stockholm 1912 45.21    NA
6      Antwerp 1920 44.69    NA
7        Paris 1924 46.16    NA
8   Amsterdam 1928 47.32 39.62
```



```
> I = !is.na(data[,3]);   men = data[I,"Men"];   women = data[I,"Women"]
> table(men>women)    # sample proportions
FALSE  TRUE
    9    14
> t.test(men[18:30],women[18:30])    # two-sample t-test
    Welch Two Sample t-test
data:  men[18:30] and women[18:30]
t = 1.1558, df = 6.5405, p-value = 0.2882
...
```

Please analyze the problem via nonparametric methods and reply Kathy's email. Use $\lesssim 200$ words possibly with tables and graphs. The report should be statistically sound and easily understandable by laymen.

1. (25%) Your report properly addresses the aforementioned problem.
2. (5%) The style of the email is appropriate. (Minor grammatical mistakes will not be penalized.)

See Remark 2 for hints.

Question 3 (★★★ — Principle and simulation study (30%)). Let $X_i = \mu_i + \varepsilon_i$, where $\mu_1, \ldots, \mu_n \in \mathbb{R}$ are some constants, and $\varepsilon_1, \ldots, \varepsilon_n$ are continuous IID RVs. Further assume that $\mu_i = g(i/(n+1))$ for some differentiable function $g : (0,1) \to \mathbb{R}$. For example, if $g(u) = u^2$ then $\mu_i = i^2/(n+1)^2$ for $i = 1, \ldots, n$. Denote the ranks of $X_1, \ldots, X_n$ by $R_1, \ldots, R_n$, respectively. The goal is to test

$$H_0 : g(\cdot) \text{ is a constant function} \qquad \text{against} \qquad H_1 : g(\cdot) \text{ is not a constant function.}$$

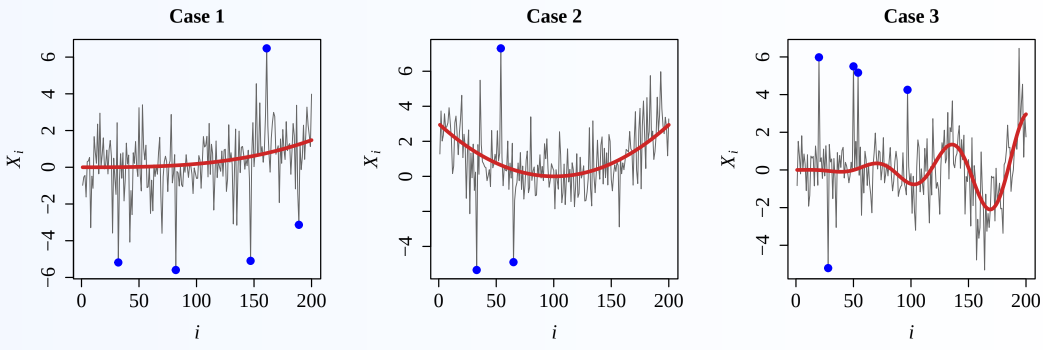
In other words, H_0 means that $\mu_1 = \cdots = \mu_n$, whereas H_1 means that not all of $\mu_1, \ldots, \mu_n$ are equal. In the lecture note, we provide a trend test for any strictly monotone function $g(\cdot)$, i.e., $\mu_1 < \mu_2 < \cdots < \mu_n$ or $\mu_1 > \mu_2 > \cdots > \mu_n$. In this question, we study a trend test for any type of non-constant function $g(\cdot)$. Define

$$\widehat{\rho}_{\mathrm{C}}^{(m)} = -2 + \frac{6}{\lambda} \sum_{i=1+m}^{n} \sum_{j=1}^{m} \min(R_i, R_{i-j})$$

for $m = 1, 2, 3, \ldots$ and a $\lambda = \lambda_{n,m} \in \mathbb{R}$ to be derived.

1. (10%) Prove that $\mathsf{E}\{\min(R_1, R_2)\} = (n+1)/3$ under H_0. Hence, prove that $\mathsf{E}(\widehat{\rho}_{\mathrm{C}}^{(m)}) = 0$ under H_0 if $\lambda = (n+1)(n-m)m$.
2. (10%) Using the λ in part 1 and the fact that $x + y - |x - y| = 2\min(x, y)$ for any $x, y \geq 0$, prove that
$$\widehat{\rho}_{\mathrm{C}}^{(1)} = \widehat{\rho}_{\mathrm{C}}(X_{1:n}, \{1, \ldots, n\}) + \frac{3}{n-1}\left(1 - \frac{R_1}{n+1} - \frac{R_n}{n+1}\right),$$
where $\widehat{\rho}_{\mathrm{C}}(x_{1:n}, y_{1:n})$ is the sample Chatterjee's correlation when the sample is $(x_1, y_1), \ldots, (x_n, y_n)$. Argue why H_0 should be rejected when $\widehat{\rho}_{\mathrm{C}}^{(1)} > c$ for some $c > 0$, and it can handle general H_1 (instead of merely monotone trend). Only a soft argument is needed for the explanation. Use $\lesssim 100$ words.
3. (10%) It is given that $\widehat{\rho}_{\mathrm{C}}^{(m)}$ behaves similarly as $\widehat{\rho}_{\mathrm{C}}^{(1)}$, except that it has a different power performance. In particular, the p-value for testing H_0 against H_1 when $\widehat{\rho}_{\mathrm{C}}^{(m)}$ is used is
$$\widehat{p}^{(m)} = 1 - \texttt{pnorm}\left(\left(\frac{2}{5mn} + \frac{8m}{15n^2}\right)^{-1/2} \widehat{\rho}_{\mathrm{C}}^{(m)}\right).$$
Let $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{IID}}{\sim} t_3$, where t_3 is a t-distribution with 3 degrees of freedom. Consider three cases:
    - (Case 1) $g(u) = \Delta u^3$;
    - (Case 2) $g(u) = \Delta 8(u - 1/2)^2$;
    - (Case 3) $g(u) = \Delta\{2u^2 \cos(6\pi u)\}$,

   where $\Delta \in [0, 1.5]$. In each case, H_0 is true when $\Delta = 0$. When $\Delta = 1.5$, the time series plots of data are shown below, where the red lines are the locations, and the blue points are the outliers.



   Consider the tests based on $\widehat{\rho}_{\mathrm{C}}^{(1)}, \widehat{\rho}_{\mathrm{C}}^{(2)}, \widehat{\rho}_{\mathrm{C}}^{(3)}$, and the trend test in Example 3.13 of the lecture note. Use simulation to compute the power curves (power against Δ) of the four tests in each of the above 3 cases when the nominal size is 5%, and $n = 200$. Briefly comment the results. Which test do you prefer? Why? (Use $\lesssim 200$ words in this part.)

See Remark 3 for hints.

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.

Remark 1 (Hints to Question 1)**.** This question provides a test for trend in scales. It is useful for testing monotone heteroscedasticity in regression models.

2. When n is an odd number, $\lceil \frac{n+1}{2} \rceil = \frac{n+1}{2}$. You may need to show that

$$\mathsf{E}(T) = \frac{n^2-1}{8} \qquad \text{and} \qquad \mathsf{Var}(T) = \frac{(n-1)(n^2+3)}{576}.$$

   You may use WolframAlpha to symbolically compute summations and simplify expressions.

3. You may use the results in the lecture note without proof, however, you must state clearly the sources, such as equation numbers and example numbers.
4. Under H_1, the test statistic t is more likely to be ... We reject ... iff ... You may have a taste on this test by using the following R code:

```r
scaleTrendTest = function(x){
    r = rank(x)
    n = length(r)
    T = sum((1:n)*abs(r/(n+1)-1/2))
    ET = (n^2-1)/8
    varT = (n-1)*(n^2+3)/576
    t = (T-ET)/sqrt(varT)
    t
}
# Under H0
n = 100
x = rnorm(n)
scaleTrendTest(x)
# Under H1
x = rnorm(n)*(n:1)/n
scaleTrendTest(x)
```

Takeaway: This question trains your ability to create a sensible rank-type test for new statistical problems.

Remark 2 (Hints to Question 2)**.**

1. You may need to be careful on the following issues:
   - avoid using the data twice,
   - appropriately handle the incomplete data,
   - understand the data, model, hypothesis, and test statistic.
2. Some remarks:
   - Is it a one-sample problem, a paired two-sample problem, an unpaired two-sample problem, a correlation problem, or something else?
   - Do you use any special methods for handling tied observations (if any)?
   - You are reminded that it is an open-ended question. Different people may have different solution.
3. A good email should
   - appropriately greet the receivers at the beginning and conclude at the end,
   - answer to all questions mentioned in the email,
   - clearly state the methods you used and the assumptions you made,
   - neatly summarize your insights obtained from your analysis, and
   - (possibly) an informative graph, list, table for describing the problem or explaining the details.

Takeaway: This question trains your ability to preprocess a dataset and analyze real-life problems.

Remark 3 (Hints to Question 3)**.**

1. Note that $\mathsf{E}\{\min(R_i, R_j)\} = \mathsf{E}\{\min(R_1, R_2)\}$ for all $i \neq j$.
2. What is the range of the "error term" $\frac{3}{n-1}(1 - \frac{R_1}{n+1} - \frac{R_n}{n+1})$?
3. Let `x` and `y` be two vectors of the same length. Then we can find out the component-by-component minimum by using the function `pmin`. Note the difference from `min`. Try the following example.

```
> x = 1:10
> y = rep(5,10)
> pmin(x,y)
 [1] 1 2 3 4 5 5 5 5 5 5
> min(x,y)
[1] 1
```

A simulated dataset in case 1 with $n = 50$ and $\Delta = 1.5$ can be generated as follows:

```
n = 50
Delta = 1.5
g1 = function(u,Delta=0){
    Delta*u^3
}
mu = g1((1:n)/(n+1), Delta=1.5)
x = mu + rt(n,3)
ts.plot(x)
points(1:n, mu, type="l", col="red")
```

Takeaway: This question trains your ability to use statistical intuition to take advantages of multiple statistics.

Evaluation. Please fill in the anonymous mid-term evaluation (click here). I value your feedback.