Homework 6
(Practice Final Exam)
STAT W4413: Nonparametric Statistics

DUE: Friday, April 29, 8:00 pm

(1) Please sign your home work with your name and UNI number.

(2) Homework must be submitted into the Statistics Homework Boxes room 904 on the 9th floor of SSW building.

(3) Homework is due Friday, April 29, 8:00 pm.

(4) No late homework, under any circumstances, will be accepted.

(5) At the end of semester, one of your lowest homework scores will be dropped before the final grade is calculated.

(6) Your submitted solutions should consist of (i) the hand written (or printout) of the results with all the details, (ii) printout of the relevant figures (if applicable), and (iii) the printout of the source code (if applicable).

(7) This homework serves as a practice final exam.

1. (15 points) Let $X \sim F$ and $Y \sim G$. Suppose that $F(x) \geq G(x)$ for all x. Prove that $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

   Solution:

$$
\begin{aligned}
\mathbb{E}(X) &= \int_0^{+\infty} x dF(x) + \int_{-\infty}^0 x dF(x) \\
&= -\int_0^{+\infty} x d(1-F(x)) + \int_{-\infty}^0 x dF(x) \\
&= -x(1-F(x))|_0^{+\infty} + \int_0^{+\infty} (1-F(x))dx + xF(x)|_{-\infty}^0 - \int_{-\infty}^0 F(x)dx \\
&= \int_0^{+\infty} (1-F(x))dx - \int_{-\infty}^0 F(x)dx \\
&\leq \int_0^{+\infty} (1-G(x))dx - \int_{-\infty}^0 G(x)dx \\
&= -x(1-G(x))|_0^{+\infty} + \int_0^{+\infty} (1-G(x))dx + xG(x)|_{-\infty}^0 - \int_{-\infty}^0 G(x)dx \\
&= -\int_0^{+\infty} x d(1-G(x)) + \int_{-\infty}^0 x dG(x) \\
&= \int_0^{+\infty} x dG(x) + \int_{-\infty}^0 x dG(x) = \mathbb{E}(Y).
\end{aligned}
$$

2. Basic results on order statistics:

   (a) (10 points) $X_1, X_2, \ldots, X_n \sim \text{Unif}(-2, 2)$ and $Y_1, Y_2, \ldots, Y_m \sim \text{Unif}(0, 2)$. Calculate the expected value of Wilcoxon rank-sum statistic.

   (b) (10 points) Let $X_1, X_2, \ldots, X_n \sim f(x)$. Derive a formula for the pdf of $X_{(1)}$.

   (c) (15 points) Let $X_1, X_2, \ldots, X_n \sim F(x)$, where F is a continuous CDF and satisfies $F(-x) = 1 - F(x)$. Calculate

   $$\mathbb{E}(rank(X_1)rank(X_1^2)),$$

   where for a function $g : \mathbb{R} \to \mathbb{R}$, $\text{rank}(g(X_1))$ denotes the rank of $g(X_1)$ between $g(X_1), \ldots, g(X_n)$.

Solution:

(a)

$$\begin{aligned}\mathbb{E}(W) &= \mathbb{E}\left(\sum_{i=1}^{n}\text{rank}(Y_i)\right)\\ &= \mathbb{E}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{1}(X_j \le Y_i) + \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{1}(Y_j \le Y_i)\right)\\ &= \mathbb{E}(U) + \frac{m(m+1)}{2}\end{aligned}$$

where $U = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{1}(X_j \le Y_i)$. Then

$$\mathbb{E}(U) = mn\text{Pr}(X_i \le Y_j) = mn\frac{3}{4}$$

Thus, $\mathbb{E}(W) = \frac{3mn}{4} + \frac{m(m+1)}{2}$.

(b) Let $F(x) = \int_{\infty}^{x} f(t)dt$ be the cdf.

$$\begin{aligned}F_{X_{(1)}}(x) &= 1 - \text{Pr}(X_{(1)} > x)\\ &= 1 - Pr(X_{(1)} > x, X_{(2)} > x, \ldots, X_{(n)} > x)\\ &= 1 - \prod_{i=1}^{n}\text{Pr}(X_i > x)\\ &= 1 - [1 - F(x)]^n\end{aligned}$$

Theb $f_{X_{(1)}}(x) = \frac{\partial}{\partial x}F_{X_{(1)}}(x) = (n-1)[1-F(x)]^{n-1}f(x)$.

(c) Since $\{X_1, X_2, \ldots, X_n\}$ and $\{-X_1, -X_2, \ldots, -X_n\}$ have the same distribution. So

$$\mu = \mathbb{E}(\text{rank}(X_1)\text{rank}(X_2^2)) = \mathbb{E}(\text{rank}(-X_1)\text{rank}((-X_2)^2)).$$

Next, note that

$$\begin{aligned}\text{rank}((-X_2)^2) &= \text{rank}(X_2^2),\\ \text{rank}(-X_1) &= n + 1 - \text{rank}(X_1).\end{aligned}$$

Hence,

$$\begin{aligned} 2\mu &= \mathbb{E}(\text{rank}(X_1)\text{rank}(X_2^2) + \text{rank}(-X_1)\text{rank}((-X_2)^2) \\ &= \mathbb{E}(\text{rank}(X_1)\text{rank}(X_2^2) + (n+1-\text{rank}(X_1))\text{rank}(-X_2^2)) \\ &= (n+1)\mathbb{E}(\text{rank}(X_2^2)) \\ &= (n+1)\frac{n+1}{2}, \end{aligned}$$

since $\text{rank}(X_i^2) \sim \text{Unif}(\{1, 2, \ldots, n\})$. Thus, $\mu = \frac{(n+1)^2}{4}$.

3. (25 points) A Poisson distribution with intensity parameter $\lambda = 3$ was proposed to model the number of arrivals per minute at a bank in New York City. Suppose that the actual arrivals per minute were observed in 200 one-minute periods over the course of a week. The results are summarized in the following table.

Arrivals	0	1	2	3	4	5	6	7	8
Observed frequency	14	31	47	41	29	21	10	5	2

Use the Pearson's χ^2 test to determine if the proposed distribution can be used to model the arrivals.

Solution

To determine whether the number of arrivals per minute follows a Poisson distribution, the null and alternative hypotheses are as follows:

- H_0: The number of arrivals per minute follows a Poisson distribution with parameter $\lambda = 3$
- H_1: The number of arrivals per minute does not follow a Poisson distribution $\lambda = 3$

The theoretical frequency for each is obtained by multiplying the appropriate Poisson probability by the sample size $n = 200$. In Matlab it corresponds to

$$200 * [poisspdf([0:7], 3), (1 - sum(poisspdf([0:7], 3)))]$$

Arrivals	0	1	2	3	4	5	6	7	8 or more
Observed frequency	14	31	47	41	29	21	10	5	2
Theoretical Frequency	9.9574	29.8722	44.8084	44.8084	33.6063	20.1638	10.0819	4.3208	2.3809

The theoretical frequency of 9 or more arrivals is 0.7606, which is less than 1.0. In order to have all categories contain a frequency of 1.0 or greater, the category 9 or more is combined with the category of 8 arrivals. So the theoretical frequency of 8 or more arrivals is $1.6203 + 0.7606 = 2.3809$. The test statistics is equal to $\chi^2 = 2.9491$ the corresponding p-value using χ^2_{9-1} distribution is 0.9375. So the decision is to accept (i.e., not to reject) H_0.

4. Explain in a few words:

   (a) (10 points) What is the main advantage of all the permutation, Monte-Carlo, and bootstrap tests and what is the main limitation that we could not resolve?
   Answer:

   - All three tests are capable of providing nonparametric tests whose distributions under H_0 do not depend on our distributional assumptions.

   (b) (5 points) What are the advantages and disadvantages of using Monte-Carlo (or bootstrap) method in the permutation test?
   Answer:

   - the limitation of the permutation test that can be resolved using Monte-Carlo (or bootstrap) test is the computational complexity. The number of permutation samples grows very rapidly as the samples of two (or $k > 2$) groups m and n increase. The computational complexity of permutation test becomes prohibitive for even small sample sizes such as $m = 20$, $n = 20$.
   - the disadvantage of Monte-Carlo test is an increased variance of the approximated p-value.

   (c) (10 points) Describe the difference between parametric and nonparametric bootstrap. What are the advantages and limitation of the two methods.
   Answer:

   - The parametric bootstrap estimates the proposed model on the original data set and simulates new samples from the estimated model. These samples are treated as a complete population and they are used to conduct inference concerning model parameters.
   - The nonparametric bootstrap treats the original data set as a complete population and draws a new, simulated sample from it, picking each observation with equal probability (allowing repeated values) and then re-running the estimation.

In result, when we have a properly specified model, simulating from the model (as in parametric bootstrap) gives more accurate results (at the same n) than does re-sampling the empirical distribution (as in nonparametric bootstrap) - parametric estimates of the distribution converge faster than the empirical distribution does. If on the other hand the parametric model is misspecified, then it is rapidly converging to the wrong distribution. This is of course just another bias-variance trade-off. If you are suspicious of your parametric modeling assumptions, choose resampling (when you can figure out how to do it, or at least until you have convinced yourself that a parametric model is very good approximation to reality).