# Final Examination

December 2019

## Instructions:

- This exam consists 6 problem-solving questions.
- The exam is open-book: hard copies of any materials are allowed.
- Show the details of your work leading to the answers.
- The marks for each question and its parts are shown in the brackets.
- Follow the Lecture Notes for any symbol that is not specified in the exam.
- Large sample approximations may be used unless otherwise specified.
- Selected critical points and formulae are provided after the end of questions.

---

## Question 1 [15 marks]

Let $X_1$ and $X_2$ be two independent continuous random variables. Define

$$S = I_{\{X_1>0\}} + 2I_{\{X_2>0\}} \quad \text{and} \quad T^+ = R_1 I_{\{X_1>0\}} + R_2 I_{\{X_2>0\}},$$

where $R_1$ and $R_2$ are the ranks of $X_1$ and $X_2$, respectively, for the Wilcoxon signed rank test.

(a) Suppose that the densities $f_1(x)$ of $X_1$ and $f_2(x)$ of $X_2$ are given by

$$f_1(x) = (1+x)I_{\{-1\le x\le 0\}} + e^{-2x}I_{\{x>0\}} \quad \text{and} \quad f_2(x) = e^{2x}I_{\{x<0\}} + (1-x)I_{\{0\le x\le 1\}}$$

Calculate $P_1 = \Pr(X_1 > 0, R_1 = 1, X_2 < 0)$ and $P_2 = \Pr(X_1 < 0, R_2 = 1, X_2 > 0)$. [6]

(b) Prove that if $X_1 \sim -X_2$ and $\Pr(X_1 > 0) = \Pr(X_1 < 0) = 0.5$, then $T^+ \sim S$, where "~" represents "identically distributed". [6]

(c) Comment on the results in parts (a) and (b). [3]

**Note:** You may use the formula $\Pr(0 < X_1 < -X_2) = \int_0^\infty F_2(-x) f_1(x)dx$

[Question 2 is on next page]

## Question 2 [15 marks]

The combined ranks of two independent random samples $X_1,\ldots,X_{11}$ and $Y_1,\ldots,Y_{10}$ are provided below:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | $Y$ | $Y$ | $X$ | $X$ | $X$ | $X$ | $X$ | $X$ | $X$ | $X$ | |
| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Sample | $X$ | $Y$ | $Y$ | $Y$ | $Y$ | $Y$ | $Y$ | $Y$ | $Y$ | $X$ | $X$ |
| Rank | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |

(a) Assume the location-scale parameter model for the two samples. Test the null hypothesis $H_0$ of no difference in location and/or dispersion parameters between the two samples by the Lepage test at appropriate level of significance. [7]

(b) It has been calculated that

$$D = \sup_{t \in \mathbb{R}} \left| F_{11}(t) - G_{10}(t) \right| = 0.6182$$

where $F_{11}(t)$ and $G_{10}(t)$ are the empirical distribution functions of the samples $X_1,\ldots,X_{11}$ and $Y_1,\ldots,Y_{10}$ respectively.

Test the general differences between the distributions of the two samples at the 5% level of significance by the two-sample Kolmogorov-Smirnov test. [4]

(c) Comment on the following issues based on the results of parts (a) and (b):

1) The overall differences between the two samples;

2) Whether the location-scale parameter model is appropriate, and why. [4]

[Question 3 is on next page]

## Question 3 [20 marks]

In a one-way layout with data $\{X_{ij}, i=1,\ldots,n_j; j=1,\ldots,5\}$, the values of

$$U_{uv} = \sum_{i=1}^{n_u}\sum_{j=1}^{n_v} I_{\{X_{iu}<X_{jv}\}} = \text{No. } \{(i,j): X_{iu} < X_{jv}, i=1,\ldots,n_u; j=1,\ldots,n_v\},\ 1 \le u < v \le 5,$$

are provided below, where $(n_1,\ldots,n_5)=(6,5,7,4,6)$:

| | $U_{uv}$ | | | |
|---|---|---|---|---|
| | $v$ | | | |
| $u$ | 2 | 3 | 4 | 5 |
| 1 | 26 | 36 | 20 | 32 |
| 2 | | 28 | 11 | 18 |
| 3 | | | 16 | 16 |
| 4 | | | | 10 |

Let $\tau_1,\ldots,\tau_5$ denote the effects of the 5 treatments and $\alpha$ the level of significance. The null hypothesis of interest is $H_0: \tau_1 = \cdots = \tau_5$.

(a) Test $H_0$ against ordered alternatives $H_1: \tau_1 \le \tau_2 \le \tau_3 \le \tau_4 \le \tau_5$ with at least one strict inequality by the Jonckheere-Terpstra test at appropriate level $\alpha$ using the large-sample normal approximation. [6]

(b) Test $H_0$ against umbrella alternatives $H_1: \tau_1 \le \tau_2 \le \tau_3 \ge \tau_4 \ge \tau_5$ with at least one strict inequality by the Mack-Wolfe test with known peak $p=3$ at appropriate level $\alpha$ using the large-sample normal approximation. [6]

(c) Test $H_0$ against umbrella alternatives by the Mack-Wolfe test with unknown peak at appropriate level $\alpha$ using the following output of R: [5]

```
> cUmbrPU(0.1, c(6,5,7,4,6))
Monte Carlo Approximation (with 10000 Iterations) used:
Group sizes: 6 5 7 4 6
For the given alpha=0.1, the upper cutoff value is Mack-Wolfe Peak Unknown
A*(p-hat)=1.9573412518, with true alpha level=0.0971
```

(d) Based on the results in parts (a) – (c), what alternatives (ordered or umbrella) have stronger support from the data according to the level of significance? Are the test results of (a) – (c) contradictive or consistent? [3]

[Question 4 is on next page]

## Question 4 [15 marks]

Consider a balanced incomplete block design (BIBD) with $k$ treatments, $n$ blocks, each treatment appearing in $p$ blocks, $s$ treatments observed in each block, and $\lambda$ pairs of treatments available in each block.

Let $c_{ij}=1$ if treatment $j$ is available in block $i$, otherwise $c_{ij}=0$, $r_{ij}$ denote the rank of $X_{ij}$ in block $i$ with $r_{ij}=0$ if $c_{ij}=0$, and $R_j=r_{1j}+\cdots+r_{nj}$.

(a) The Durbin-Skillings-Mack test statistic $D$ for BIBD is defined by

$$D=\frac{12}{\lambda k(s+1)}\sum_{j=1}^{k}\left(R_j-\frac{p(s+1)}{2}\right)^2=\frac{12}{\lambda k(s+1)}\sum_{j=1}^{k}R_j^2-\frac{3(s+1)p^2}{\lambda}$$

Find the formulae of $\mathrm{E}[R_j]$ and $\mathrm{Var}(R_j)$ using

$$\mathrm{E}[r_{ij}]=\frac{s+1}{2}I_{\{c_{ij}=1\}},\quad \mathrm{Var}(r_{ij})=\frac{(s+1)(s-1)}{12}I_{\{c_{ij}=1\}},$$

and the independence of $\{r_{ij}\}$ between $i=1,\ldots,n$. Then prove $\mathrm{E}[D]=k-1$. [5]

(b) The following table presents the data $\{X_{ij}\}$ in an incomplete block design:

| Block | Treatment | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 21 | 15 | 17 | – | 28 |
| 2 | 25 | – | 19 | 35 | 32 |
| 3 | 39 | 32 | 35 | 44 | – |
| 4 | – | 22 | 16 | 24 | 30 |
| 5 | 38 | 34 | – | 45 | 42 |

Test the null hypothesis $H_0:\tau_1=\cdots=\tau_k$ against general alternatives at appropriate level of significance. [6]

(c) Given that $q_{0.1}=3.479$ for $k=5$, decide the differences between treatment effects $\tau_1,\ldots,\tau_k$ based on the data in part (b) by the Skillings-Mack two-sided all-treatment multiple comparison procedure for BIBD with $\alpha=0.1$. [4]

[Question 5 is on next page]

## Question 5 [20 marks]

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be i.i.d. pairs of continuous random variables. Define

$$K=\sum_{1\le u<v\le n} Q_{uv} \quad \text{with} \quad Q_{uv}=Q_{uv}=\begin{cases} 1 & \text{if } (X_u-X_v)(Y_u-Y_v)>0 \\ 0 & \text{if } (X_u-X_v)(Y_u-Y_v)=0, \quad 1\le u<v\le n. \\ -1 & \text{if } (X_u-X_v)(Y_u-Y_v)<0 \end{cases}$$

The following data are observed from $(X_1,Y_1),\ldots,(X_n,Y_n)$:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 4 | 7 | 12 | 12 | 12 | 21 | 26 | 35 | 35 | 64 |
| $Y_i$ | 11 | 8 | 15 | 12 | 33 | 24 | 16 | 19 | 54 | 48 |

where $X_1,\ldots,X_n$ are rearranged in nondecreasing order. Let $g$ denote the number of tied groups among $X_1,\ldots,X_n$ and $t_i$ the size of group $i$, $i=1,\ldots,g$.

(a) Let $\tau$ denote the Kendall correlation coefficient. Test the null hypothesis $H_0$ of independent $(X_i,Y_i)$ against $H_1:\tau>0$ at the 1% level of significance based on $K$ and the following null variance of $K$ with ties in $X_1,\ldots,X_n$:

$$\operatorname{Var}_0(K)=\frac{n(n-1)(2n+5)}{18}-\frac{1}{18}\sum_{i=1}^{g}t_i(t_i-1)(2t_i+5) \qquad [5]$$

(b) You are given $C_1=C_2=C_3=C_4=7$, $C_5=5$, $C_6=4$, $C_7=3$, $C_8=1$, $C_9=7$, where $C_i=\sum_{t\neq i}Q_{it}$, $i=1,\ldots,n$. Find an approximate 95% confidence interval of $\tau$ (you may use the formula with no ties). [5]

(c) Let $r$ represent the correlation coefficient of $(X_i,Y_i)$. Test the null hypothesis $H_0$ of independent $(X_i,Y_i)$ against $H_1:r>0$ at the 1% level of significance based on the Spearman rank correlation coefficient $r_s$ adjusted for ties:

$$r_s=\frac{1}{\sqrt{n(n^2-1)-A}\sqrt{n(n^2-1)}}\left[n(n^2-1)-6\sum_{i=1}^{n}(R_i-S_i)^2-\frac{A}{2}\right],$$

where $(R_i,S_i)$ are the ranks of $(X_i,Y_i)$ and $A=t_1(t_1^2-1)+\cdots+t_g(t_g^2-1)$. Comment on the relationship between $(X_i,Y_i)$. [5]

(d) Define $\delta=\Pr(Q_{12}=1,Q_{13}=1)$. Given $\mathrm{E}[Q_{is}Q_{it}]=4\delta-1-2\tau$ for $1\le s\neq t\neq i\le n$, derive an unbiased estimator of $\delta$ (you are not required to calculate it) based on $K$ and $T=\sum_{i=1}^{n}\sum_{1\le s\neq t\neq i\le n}Q_{is}Q_{it}$. [5]

[Question 6 is on next page]

**Question 6** [15 marks]

The table below presents the data of a regression line with slope $\beta$:

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $Y_i$ | 6.5 | 7.8 | 6.2 | 3.6 | 4.3 | 3.2 | 3.5 |

(a) Estimate the slope $\beta$ and intercept $\alpha$ of the regression line by the nonparametric method related to the Theil test. [6]

(b) Let $k_\alpha$ denote the critical point such that $\Pr(K \geq Nk_\alpha) = \alpha$, where $K$ represents the Kendall statistic from $n$ pairs of random variables and $N = n(n-1)/2$.
Given $k_{0.0345} = 0.619$, find an exact confidence interval of the slope $\beta$ at the appropriate confidence level. [4]

(c) Choose an appropriate one-sided alternative: $\beta > 0$ or $\beta < 0$, based on the results in parts (a) and (b). Then determine whether there is sufficient evidence to support the alternative you have chosen by the Theil test with the normal approximation at the 2.5% level of significance. [5]

[End of questions]

## Selected critical points:

$\Pr(Z \geq z_\alpha) = \alpha$ for $Z \sim N(0,1)$ and $\Pr\left(X \geq \chi^2_{k,\alpha}\right) = \alpha$ for $X \sim \chi^2_k$

| $\alpha$ | 0.001 | 0.005 | 0.01 | 0.025 | 0.05 | 0.10 | 0.25 | 0.40 |
|---|---|---|---|---|---|---|---|---|
| $z_\alpha$ | 3.090 | 2.576 | 2.326 | 1.960 | 1.645 | 1.282 | 0.674 | 0.253 |
| $\chi^2_{1,\alpha}$ | 10.83 | 7.879 | 6.635 | 5.024 | 3.841 | 2.706 | 1.323 | 0.708 |
| $\chi^2_{2,\alpha}$ | 13.82 | 10.60 | 9.210 | 7.378 | 5.991 | 4.605 | 2.773 | 1.833 |
| $\chi^2_{3,\alpha}$ | 16.27 | 12.84 | 11.34 | 9.348 | 7.815 | 6.251 | 4.108 | 2.946 |
| $\chi^2_{4,\alpha}$ | 18.47 | 14.86 | 13.28 | 11.14 | 9.488 | 7.779 | 5.385 | 4.045 |

Selected formulae are provided on next two pages.

## Selected formulae:

1. Mean and variance of the Wilcoxon rank sum statistic $W$ under $H_0$:

$$\mathrm{E}_0[W]=\frac{n(N+1)}{2},\quad \mathrm{Var}_0(W)=\frac{mn(N+1)}{12}$$

2. Mean and variance of the Ansari-Bradley test statistic $C$ under $H_0$:

For even $N$: $$\mathrm{E}_0[C]=\frac{n(N+2)}{4},\quad \mathrm{Var}_0(C)=\frac{mn(N+2)(N-2)}{48(N-1)}$$

For odd $N$: $$\mathrm{E}_0[C]=\frac{n(N+1)^2}{4N},\quad \mathrm{Var}_0(C)=\frac{mn(N+1)(N^2+3)}{48N^2}$$

3. The two-sample Kolmogorov-Smirnov test:

$$J=\frac{mn}{d}D=\frac{mn}{d}\sup_{t\in\mathbb{R}}\left|F_m(t)-G_n(t)\right|,\quad J^*=\frac{J}{\sqrt{mnN}},\quad \Pr\left(J^*\geq x\right)\approx 2e^{-2x^2}$$

4. Mean and variance of the Jonckheere-Terpstra test statistic $J$ under $H_0$:

$$\mathrm{E}_0[J]=\frac{1}{4}\left(N^2-\sum_{i=1}^{k}n_i^2\right),\quad \mathrm{Var}_0(J)=\frac{1}{72}\left[2N^3+3N^2-\sum_{i=1}^{k}n_i^2(2n_i+3)\right]$$

5. Mean and variance of the Mack-Wolf statistic $A_p$ under $H_0$:

$$\mathrm{E}_0[A_p]=\frac{1}{4}\left(N_1^2+N_2^2-\sum_{i=1}^{k}n_i^2-n_p^2\right)$$

$$\mathrm{Var}_0(A_p)=\frac{1}{72}\left[2(N_1^3+N_2^3)+3(N_1^2+N_2^2)-\sum_{i=1}^{k}n_i^2(2n_i+3)-n_p^2(2n_p+3)\right]+\frac{1}{6}\left(n_pN_1N_2-n_p^2N\right)$$

6. Mean and variance of $U_{\bullet q}$ under $H_0$:

$$\mathrm{E}_0[U_{\bullet q}]=\frac{n_q(N-n_q)}{2},\quad \mathrm{Var}_0(U_{\bullet q})=\frac{n_q(N-n_q)(N+1)}{12}$$

7. Skillings-Mack multiple two-sided all-treatment comparison procedure for BIBD:

$$\text{Decide } \tau_u\neq\tau_v \text{ if } \left|R_u-R_v\right|\geq q_\alpha\sqrt{\frac{(s+1)(ps-p+\lambda)}{12}}$$

8. An estimate of $\mathrm{Var}(\hat{\tau})$ for the estimator $\hat{\tau}$ of the Kendall correlation coefficient:

$$\hat{\sigma}^2 = \frac{2}{n(n-1)}\left[\frac{2(n-2)}{n(n-1)^2}\sum_{i=1}^{n}\left(C_i - \bar{C}\right)^2 + 1 - \hat{\tau}^2\right], \quad C_i = \sum_{i \neq t}^{n} Q_{it}\,, \quad \bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

9. Theil statistic for the slope of a regression line:

$$C = \sum_{i<j}^{n} c\left(D_j - D_i\right), \quad D_i = Y_i - \beta_0 x_i\,, \quad c(a) = I_{\{a>0\}} - I_{\{a<0\}}$$

with

$$\mathrm{E}_0[C] = 0 \quad \text{and} \quad \mathrm{Var}_0(C) = \frac{n(n-1)(2n+5)}{18}$$

10. Estimate of the slope $\beta$ of a regression line related to the Theil test:

$$\hat{\beta} = \mathrm{median}\left\{S_{ij}, 1 \le i < j \le n\right\}, \quad \text{where } S_{ij} = \frac{Y_j - Y_i}{x_j - x_i}$$

11. Estimate of the intercept $\alpha$ of a regression line related to the Theil test:

$$\hat{\alpha} = \mathrm{median}\{A_1, \ldots, A_n\}, \quad \text{where } A_i = Y_i - \hat{\beta} x_i\,, \; i = 1, \ldots, n\,.$$

12. Confidence interval of slope:

$$\left(S_{(M)}, S_{(Q+1)}\right) \text{ with } M = \frac{N - C_\alpha}{2} \text{ and } Q = \frac{N + C_\alpha}{2} = M + C_\alpha\,,$$

where $N = n(n-1)/2$ and $C_\alpha = N k_{\alpha/2} - 2$.