Nonparametric Statistics (STAT 3504): Exam 2, Fall 2017

Instructor: Robert B. Gramacy

Instructions

1. Please direct all questions concerning this exam only to the instructor. Do not collaborate with a student in the classroom during the exam.
2. Please turn off (not on vibrate) all cellular phones before starting this exam.
3. When you finish your exam, print your name at the bottom of this cover sheet and sign the honor pledge.
4. You are allowed one sheet of notes, front and back, that you produced yourself, a calculator and a writing device.
5. For all calculations on probability distributions, consult the table on the sheet provided.
6. For two-sided tests please calculate the p-value as $\varphi \approx 2 \times \min\{\mathbb{P}(T \leq t), \mathbb{P}(T \geq t)\}$.
7. All hypothesis tests default to an $\alpha = 0.05$ level unless otherwise stated; CIs may specify a different α.

I agree to take the exam under the above directions, and have completed this exam in compliance with the Virginia Tech Honor Code.

Name: Key

Signature: ______________________

Date: ______________________

Problem 1 [20 pts]

greater than

Four volunteers were assigned to each of two weight-reducing plans. The assignment of the volunteers to the plans was at random, and it was assumed that the 12 volunteers are a random sample of people who might try a weight-reducing program. Determine if the mean amount of weight lost from Plan ~~A~~ B is ~~less than~~ from Plan ~~B~~ A.

Plan A	Plan B
15	2
19	12
14	14
6	4

a. (2 pts) Which test are you using?

Mann-Whitney / Wilcox test for heterogeneity in means

upper ~~lower~~ tail

b. (3 pts) State the null and alternative hypotheses.

H_0: Distribution of A & B same except perhaps $E(A) > E(B)$

$[F_A(x) \le F_B(x)]$

H_1: $E(A) < E(B)$ [or equiv $F_A(x) > F_B(x)$]

Students may also write $H_0: \mu_A \geq \mu_B$

$H_1: \mu_A < \mu_B$

as long as they say what μ means

c. (6 pts) Compute the test statistic.

(x) Rank A = 7, 8, 6, 4

$T = \sum_{i=1}^{4} R(X_i) = 25 \qquad T' = T - \frac{n_x(n_x+1)}{2} = \frac{25 - 4(5)}{2}$

$T' = 15$

There are no ties, so no need to calculate T_1 via CLT.

But in case $T_1 = 2.020726$

d. (6 pts) Calculate the p-value or rejection region.

$T \sim \text{MannWitney}(4,4)$

$p = P(T \geq t)$ or $p = P(T' \geq t')$

$= 1 - P(T \leq t-1)$ for $T' \sim \text{Wilcox}(4,4)$

$= 1 - 0.9714$ $= 1 - P(T' \leq t'-1)$

$= 0.0286$ $= 1 -$

e. (3 pts) What do you conclude?

reject H_0 @ 5% level and conclude that the (mean) amount of weight loss from Plan A exceeds Plan B

Problem 2: [20 pts]

Consider the two weight-reduction plans above, and a third Plan C. Researchers believe that a difference in the median weight lost could be interpreted as a difference in the value of the plan. Perform a test to determine if the three plans have the same median.

Plan A	Plan B	Plan C
15	2	19
19	12	7
14	5	32
6	4	20

a. (2 pts) Which test are you using?

The median test

b. (3 pts) State the null and alternative hypotheses.

H_0: All three plans have the same median

H_1: At least two have different medians

c. (6 pts) Compute the test statistic.

First compute the grand median. The sorted values are 2, 4, 5, 6, 7, 12, 14, 15, 19, 19, 20, 32

→ med is 13

So the table is

	A	B	C	tot
>13	3	0	3	6
≤13	1	4	1	6
tot	4	4	4	12

$$T = \left(\sum O_{1i}^2 / n_i\right) \frac{N^2}{ab} - \frac{Na}{b}$$

$$= 6$$

d. (6 pts) Calculate the p-value or rejection region.

$$T \sim \chi^2_{c-1} \equiv \chi^2_2 \quad \text{and} \quad p = P(T > t)$$
$$= 1 - P(T \leq 6)$$
$$= 1 - 0.9502129$$
$$= 0.049787$$

e. (3 pts) What do you conclude?

Reject H_0 (barely) at the 5% level and conclude that not all plans have the same median

Problem 3: [20 pts]

The number of babies born in a certain hospital last year was as follows. In the Winter there were 63 babies born, in the Spring 54, in the Summer 24 and in the Fall 55. Test the hypothesis that the number of births are uniformly distributed over the four seasons of the year, i.e, there is equal probability to be born in each of the seasons.

a. (2 pts) Which test are you using?

χ^2 goodness of fit test

b. (3 pts) State the null and alternative hypotheses.

H_0: $p_F = p_W = p_{Sp} = p_{Su} = 1/4$ where p_{seas} is the probability that a baby is born in that season

H_1: At least one of the probabilities differs from another. (and thus differs from 1/4)

c. (6 pts) Compute the test statistic.

N = 196 total observations, therefore we have the following observed & expected counts

O: 63 54 24 55

E: 49 49 49 49

so $T = \sum_{j=1}^{4} \frac{(O_j - E_j)^2}{E_j} = 18$

d. (6 pts) Calculate the p-value or rejection region.

$T \sim \chi^2_{c-1} \equiv \chi^2_3$ & $p = P(T > t)$

$= 1 - P(T < 18)$

$= 1 - 0.9995602$

$= 0.0004398$

e. (3 pts) What do you conclude?

Reject H_0 & conclude that there is a difference in birth rate over seasons

Problem 4: [20 pts]

Is there significant correlation between the age at which a U.S. president was inaugurated for the first time and the age at which he died?

President	Inaugurated	$R(I)$	Died	$R(D)$
Jefferson	57	7	83	7
Var Buren	54	3.5	79	5
Pierce	48	2	64	3
Grant	46	1	63	2
McKinley	54	3.5	58	1
Wilson	56	5.5	67	4
Nixon	56	5.5	81	6
$\sum_i R_i^2$		139		140
$\sum_i R_i^I R_i^D$		133		

a. (2 pts) Which test are you using?

Spearman ρ test for association
(correlation or indep)

b. (3 pts) State the null and alternative hypotheses.

two-tailed test:

H_0: Inaguration & Death year mutually independent
H_1: Either (a) there is a tendency for older inaugurations to result in early death, or (b) older inaugurations in later death

c. (6 pts) Compute the test statistic. $X \equiv I$ & $Y \equiv D$

$$\hat{\rho} = \frac{\sum_{i=1}^{7} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{7} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{7} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2}} = \frac{133-112}{\sqrt{133-112}\sqrt{140-112}}$$

$$= 0.7637$$

but there are many ties in the ranks, so we prefer

$$z = \hat{\rho}\sqrt{n-1}$$ (for full credit)

d. (6 pts) Calculate the p-value or rejection region.

$\hat{\rho} \sim \text{Spearman}(7)$

$$\psi = 2\min\left(P(R \le \hat{\rho}), P(R > \hat{\rho})\right)$$
$$= 2\min\left(P(R \le 0.76), 1 - P(\text{''})\right)$$
$$= 2\min\left(0.97599, 0.02400\right)$$
$$= 0.0480$$

but CLT gives

$$\psi = 2 \cdot P\left(Z \le -|\hat{\rho}|\sqrt{n-1}\right)$$
$$= 2 \cdot P(Z \le -1.8706)$$
$$= 2 \cdot 0.0307 = 0.06140$$

(full credit)

e. (3 pts) What do you conclude?

Fail to reject H_0 & conclude there is no correlation between Inaug & Death year

(partial credit)

Reject H_0.

Problem 6: [20 pts]

The following table refers to a retrospective study of lung cancer and tobacco smoking among patients in English hospitals. The table compares lung cancer patients with control patients having other diseases, according to the number of cigarettes smoked daily over a 10-year period preceding the onset of disease.

Daily Avg of Cigarettes	Lung Cancer Patients	Control Patients	tot
< 5	26 (67.5)	109 (67.5)	135
5-24	946 (978)	1010 (978)	1956
25+	313 (239)	166 (239)	479
tot	1285	1285	2570

Compute the following:

(10) a. ~~(5 pts)~~ Calculate Cramèr's contingency coefficient C_{cc}, and report its range.

$$E_{ij} = (1) = \frac{n_i c_j}{N} \qquad T = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 98.23$$

$$R_1 = \frac{T}{N(\min\{r,c\}-1)} = \frac{98.23}{N\cdot(2-1)} = 0.03822$$

$$C_{cc} = \sqrt{R_1} = 0.1955 \text{ w/ range } [0,1]$$

(5) b. ~~(2.5 pts)~~ Pearson's Contingency Coefficient, R_2, and report its range.

$$R_2 = \sqrt{\frac{T}{N+T}} = \sqrt{\frac{98.23}{2570+98.23}} = 0.19187 \text{ with range } \left[0, \sqrt{q-1/q}\right] = [0, 0.707]$$

(5) c. ~~(2.5 pts)~~ Pearson's Mean-Square Contingency Coefficient, R_3, and report its range.

$$R_3 = \frac{T}{N} = \frac{98.23}{2570} = 0.0382 \text{ w/ range } [0, q-1] = [0,1]$$