Chapter14 Nonparametric Statistics

Introduction:

________________________: methods for making inferences about population parameters (confidence interval and hypothesis testing) rely on the assumptions about probability distribution of sampled population. (t-test, ANOVA)

_________________________: statistical tests that do not rely on any underlying assumptions about the probability distribution of the sampled population.

_______________: the branch of inferential statistics devoted to distribution-free tests.

_________________________: nonparametric tests (or statistics) based on the ranks of measurements.

14.3 comparing two populations: based on independent samples

When t-test is not appropriate, we can use nonparametric methods to compare two populations.

________________________: comparing the location of center for two or more non-normal populations. It tests the hypothesis that the **probability distributions associated with the two populations are equivalent.**

- **Wilcoxon rank sum test:**

Define: D_1 : the ______________________for population 1,

D_2 : the ______________________for population 2.

Three relationship between D_1 and D_2 :

1. Hypothesis:

H_0 : _________________________________

H_a : _________________________________

(H_a : D_1 is shifted to the ___________of D_2)

(H_a : D_1 is shifted to the ___________of D_2)

2. Test statistic:

Step 1. ________ all the data values (both samples) from 1 to N, ($N = n_1 + n_2,\ \ n_1 < n_2$)

(Note: assign the ___________ of the ranks to each of the tied observations)

Step 2. **Test statistic** T_1 = Sum the ranks for the _______________(n_1).

3. Rejection region:_________________ if $H_a : D_1$ and D_2 are not identical;

________ if $H_a : D_1$ is shifted to the left of D_2;

________ if $H_a : D_1$ is shifted to the right of D_2.

Critical values T_L *and* T_U are from Table XII, p807.

TABLE XII Critical Values of T_L and T_U for the Wilcoxon Rank Sum Test: Independent Samples

Test statistic is the rank sum associated with the smaller sample (if equal sample sizes, either rank sum can be used).

a. α = .025 one-tailed; α = .05 two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

b. α = .05 one-tailed; α = .10 two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

Source: From F. Wilcoxon and R. A. Wilcox, "Some Rapid Approximate Statistical Procedures," 1964, 20–23.

- **Condition required for a valid rank sum test:**

1. the two samples are_________________________________;
2. the two probability distributions from which the samples are drawn are _________.

Example1: DRUGS:

At $\alpha = 0.05$, Do the data provide sufficient evidence to indicate a shift in the probability distributions for drug A and B?

Drug A	Drug B
Reaction time(seconds)	Reaction time(seconds)
1.96	2.11
2.24	2.41
1.71	2.07
2.41	2.71
1.62	2.50
1.93	2.84
	2.88

Define: D_1 : the probability distribution for reaction times under drug A**(__________)**

D_2 : the probability distribution for reaction time under drug B;

H_0 : ________________________________

H_a :_________________________________

Example 2: VENDERS: Raw material for an industrial process is provided by two different venders. Four batches of raw material are purchased from each vender, and the strength of the finished product (in p.s.i) is measured for each batch.

At $\alpha = 0.05$, do the data provide sufficient evidence to indicate the probability distributions of product strength for vender 1 is shifted to left of the vender 2?

Vender 1	Vender 2
874	890
883	908
900	924
915	930

The Wilcoxon Rank sum test for large sample (____________________):

Hypothesis:

$H_0 : D_1$ and D_2 are identical,

$H_a : D_1$ and D_2 are not identical, R.R: ___________

($H_a : D_1$ is shifted to the left of D_2, R.R: __________)

($H_a : D_1$ is shifted to the right of D_2, R.R: __________)

Test statistic: $$z = \frac{T_1 - \dfrac{n_1(n_1+n_2+1)}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1+n_2+1)}{12}}}$$

Example: Verbal SAT scores for students randomly selected from two different schools are listed below. Use the Wilcoxon rank sum procedure to test the claim that there is no difference in the scores from each school. Use $\alpha = 0.05$.

School 1	550, 520, 770, 480, 750, 530, 580, 780, 610, 590, 730, 750
School 2	490, 440, 680, 430, 710, 590, 690, 550, 530, 630, 640, 540

14.5 comparing three or more populations: completely randomized design

Nonparametric method also can used to analyze the data from CRD.

__________________________________: no assumptions concerning the population probability distribution;

- **Rank procedure:**

1. Rank _______ observations from smallest (1) to largest (n),

2. Add up the ranks for___________________, R_j, $j=1,\ldots,k$

- **Hypothesis:**

H_0: ____

H_a: ____

Test statistic: _________________________________

Rejection region: ___________, with ______________degree of freedom. (p798)

- **Conditions required for the valid Kruskal-wallis test:**

1. The k samples are_________________________,

2. there are _____________________measurements in each sample;

3. the k probability distributions from which the samples are drawn are___________.

Example1: OZONE EFFECT: To study the effects of ozone exposure on airway resistance, 15 healthy male subjects were randomly assigned to each of 3 exposure levels: 0.1ppm, 0.6ppm, 1.0ppm for one hour**.** The differences of airway resistance after exposure and before exposure for each subject were recorded. At α =0.01, is there evidence to conclude there exists a difference in the probability distribution among three ozone exposure levels?

0.1 ppm (1)	0.6 ppm (2)	1.0 ppm (3)
-0.08	-0.11	0.06
0.14	0.01	0.14
0.21	0.07	0.19
0.30	0.14	0.20
0.50	0.18	0.34

H_0: ____

H_a: ____

Example2: Vehicle miles:

The U.S federal highway administration conducts annual survey on motor vehicle travel by type of vehicle and publishes in highway statistics. Independent random samples of cars, buses, and trucks provided the data on number of miles driven last year, in thousands. At $\alpha = 0.05$, do the data provide sufficient evidence to conclude that a **difference exists** in the probability distribution of last year's miles among cars, buses, and trucks?

cars	buses	trucks
15.3	1.8	24.6
15.3	7.2	37.0
2.2	7.2	21.2
6.8	6.5	23.6
34.2	13.3	23.0
8.3	25.4	15.3
12.0		57.1
7.0		14.5
9.5		26.0
1.1		

H_0:

H_a:

14.6 Comparing three or more populations: randomized Block design

______________________: no assumptions concerning the population probability distribution;

- **Rank procedure:**

1. Rank observations within_________________;

2. Add up the ranks for__________________; $R_j, \quad j=1,...,k$

- **Hypothesis:**

H_0: ____

H_a: ____

- Test statistic: ____

____ : # of blocks; _____: # of treatments.

Rejection region: _____________, with ___________degrees of freedom.(p798)

- **Conditions required for a valid Friedman test:**

1. the treatments are ____________ assigned to experimental units within the blocks;
2. the measurements can be ranked within__________;
3. the k probability distributions from which the samples within each block are drawn are____________________.

Example1: Reaction2,

(when a drug effect is short-lived， no carryover effect; and the drug effect varies greatly from person to person, it may be useful to employ a RBD to find the effect of a drug.)

At $\alpha = 0.05$, do the data provide sufficient evidence to conclude the probability distributions of the reaction times for the three drugs differ in location?

Reaction time for three drugs:

subject	Drug A	Drug B	Drug C
1	1.21	1.48	1.56
2	1.63	1.85	2.01
3	1.42	2.06	1.70
4	2.43	1.98	2.64
5	1.16	1.27	1.48
6	1.94	2.44	2.81

Example2. In a study comparing the effects of four energy drinks on running speed, seven runners were timed (in seconds) running four miles. On each day, they were given a single energy drink. The data are listed below. Is there evidence of a difference in the probability distributions of the running times among the four drinks? Use $\alpha = 0.10$.

	drink			
runner	1	2	3	4
1	1226	1226	1273	1244
2	1129	1035	1151	1159
3	1229	1357	1291	1303
4	1256	1217	1272	1267
5	1159	1121	1215	1220
6	1318	1295	1348	1344
7	1220	1261	1257	1243

14.7 Rank Correlation

In chapter 11, we have introduced a correlation coefficient r based on the observations of two numerical variables.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

x	y	
x_1	y_1	
x_2	y_2	
….	…..	
x_n	y_n	

Here we introduce a correlation coefficient based on ____________ of observations:

- **_________________________________:** it provides a measure of correlation between ranks.

$$r_s = \frac{SS_{uv}}{\sqrt{SS_{uu}SS_{vv}}}$$

Where

$u_i =$ _______ of the ith observation in sample 1

$v_i =$ ________of the ith observation in sample 2

$n =$ _________________________

$$SS_{uv} = \sum (u_i - \overline{u})(v_i - \overline{v}) = \sum u_i v_i - \frac{(\sum u_i)(\sum v_i)}{n}$$

$$SS_{uu} = \sum (u_i - \overline{u})^2 = \sum u_i^2 - \frac{(\sum u_i)^2}{n}$$

$$SS_{vv} = \sum (v_i - \overline{v})^2 = \sum v_i^2 - \frac{(\sum v_i)^2}{n}$$

- **Shortcut formula for r_s : (no ties or less ties relative to n)**

$$r_s =$$

Where ________________ (rank difference of ith observations between sample 1 and sample 2)

Note: 1. The value of r_s always falls between______________,

2. ________indicates perfect positive correlation and ________ indicates perfect negative correlation,

3. The closer r_s falls to +1 or -1, the ________the correlation between the ranks; the closer r_s is to 0, the __________the correlation.

- **Spearman's rank correlation test:**

Define ρ: ____

$H_0 : \rho = 0$ (____)

$H_a : \rho \neq 0$ (____)

Test statistic: ____

Rejection region:_______________ if $H_a : \rho \neq 0$;

______________ if $H_a : \rho < 0$

_______________ if $H_a : \rho > 0$

Critical value: $r_{s,\alpha}$ from Table XIV, p809.

TABLE XIV Critical Values of Spearman's Rank Correlation Coefficient

The α values correspond to a one-tailed test of H_0: $\rho = 0$. The value should be doubled for two-tailed tests.

n	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$	n	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$
5	.900	—	—	—	**18**	.399	.476	.564	.625
6	.829	.886	.943	—	**19**	.388	.462	.549	.608
7	.714	.786	.893	—	**20**	.377	.450	.534	.591
8	.643	.738	.833	.881	**21**	.368	.438	.521	.576
9	.600	.683	.783	.833	**22**	.359	.428	.508	.562
10	.564	.648	.745	.794	**23**	.351	.418	.496	.549
11	.523	.623	.736	.818	**24**	.343	.409	.485	.537
12	.497	.591	.703	.780	**25**	.336	.400	.475	.526
13	.475	.566	.673	.745	**26**	.329	.392	.465	.515
14	.457	.545	.646	.716	**27**	.323	.385	.456	.505
15	.441	.525	.623	.689	**28**	.317	.377	.448	.496
16	.425	.507	.601	.666	**29**	.311	.370	.440	.487
17	.412	.490	.582	.645	**30**	.305	.364	.432	.478

Source: From E. G. Olds, "Distribution of Sums of Squares of Rank Differences for Small Samples," *Annals of Mathematical Statistics*, 1938, 9.

Note: The α values correspond to a one-tailed test of $H_0 : \rho = 0$.

- **Condition required for a valid Spearman's Test:**

1. The sample of experimental units on which the two variables are measured is _________________selected,
2. The probability distributions of the two variables are________________.

Example: The number of absences and the final grades of 9 randomly selected students from a statistics class are given below. Can you conclude that there is a correlation between the final grade and the number of absences? Use $\alpha = 0.05$.

student	absence	Final grade		
1	0	98		
2	3	86		
3	6	80		
4	4	82		
5	9	71		
6	2	92		
7	15	55		
8	8	76		
9	5	82		