

BIOS 600: Principles of Statistical Inference

Nonparametric Methods

Fall 2012

Reading

- ▶ Pagano and Gauvreau, Chapter 13

Why Nonparametric Methods?

For the methods we have studied so far, we have assumed the populations from which the data were drawn were either normally distributed or approximately so. This is a necessary property for the tests to be valid. Because the distributional form is assumed known, with only values of μ and σ unknown, these methods are known as *parametric* methods.

Nonparametric methods make fewer assumptions about the nature of the underlying distributions and may be appropriate when some assumptions of *parametric* methods are not satisfied.

Sign Test

The *sign test* is a nonparametric alternative to the paired t-test. It does not assume normality but just requires that observations are independent. The null hypothesis of the sign test is that the median difference among pairs in the underlying population is 0.

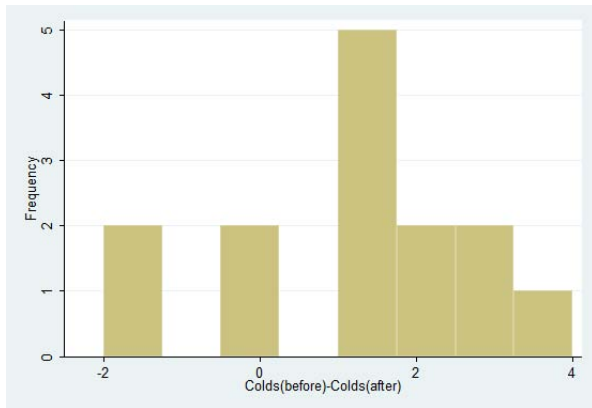
Sign Test: Cold Data

A medical researcher claims that a new vaccine will decrease the number of colds in adults. You randomly select 14 adults and record the number of colds each has in a one-year period. After giving the vaccine to each adult, you again record the number of colds each has in a one-year period. At $\alpha = 0.05$, do the data support the researchers claim?

<i>Year</i>	<i>Number of Colds</i>													
Before vaccine	3	4	2	1	3	6	4	5	2	0	2	5	3	3
After vaccine	2	1	0	3	1	3	3	2	2	2	3	4	3	2

What is H_0 ?

Sign Test: Cold Data



Sign Test

The sign test has just a few simple steps.

- ▶ Take the difference of each pair of observations (e.g., before-after)
- ▶ Record the sign of each difference (+, −, or 0)
- ▶ Count the number of + signs
- ▶ Test using the Bernoulli distribution with n =number of pairs with nonzero differences and $p = 0.50$, as we would expect the same number of positive and negative signs under H_0

Sign Test for Cold Data

Take note: these differences do NOT look very normal!

<i>Year</i>	<i>Number of Colds</i>													
Before vaccine	3	4	2	1	3	6	4	5	2	0	2	5	3	3
After vaccine	2	1	0	3	1	3	3	2	2	2	3	4	3	2
Difference	1	3	2	-2	2	3	1	3	0	-2	-1	1	0	1
Sign	+	+	+	-	+	+	+	+	0	-	-	+	0	+

$n = 12$ nonzero differences, # of + signs: 9

```
. bitesti 12 9 .5
```

N	Observed k	Expected k	Assumed p	Observed p
12	9	6	0.50000	0.75000
Pr(k >= 9) = 0.072998 (one-sided test)				
Pr(k <= 9) = 0.980713 (one-sided test)				
Pr(k <= 3 or k >= 9) = 0.145996 (two-sided test)				

We fail to reject H_0 and conclude there is no difference in the median # of colds in the years before and after vaccine administration.

Wilcoxon Signed-Rank Test

- ▶ Sign test is appealing because it avoids distributional assumptions
- ▶ However, it ignores the magnitude of the differences
- ▶ If we are willing to assume something about the distribution (that the differences are symmetrically distributed around the median), we can incorporate the magnitude of the differences and gain considerable power using the *Wilcoxon Signed-Rank Test*
- ▶ Wilcoxon signed-rank test is an alternative to the one-sample or paired t-test
- ▶ H_0 : the median difference in the underlying population is 0

Rank-Based Methods

Many nonparametric methods have a similar flavor to parametric methods but are computed not based on the observed data but on the ranks.

The *rank* of an observation, among a set of observations, is its position when the observations are ordered from smallest to largest. The smallest observation has rank 1, next smallest has rank 2, and so forth. If observations are tied, the rank assigned to each is the average of the ranks appropriate to the equal numbers.

Why Ranks?

Using the ranks provides robustness against outliers. Ranking is fairly simple though care must be taken in the presence of ties. Consider the following body weights of adult males and their ranks.

- ▶ Weights: 175, 169, 190
- ▶ Ranks: 2, 1, 3
- ▶ Weights: 175, 169, 380
- ▶ Ranks: 2, 1, 3 (robust to outlier)
- ▶ Weights: 175, 175, 190
- ▶ Ranks: 1.5, 1.5, 3 (take sum of ranks and divide by number tied $\frac{1+2}{2} = 1.5$)

Wilcoxon Signed-Rank Test

Like the sign test, the Wilcoxon Signed-Rank test is for paired differences (or a single sample) and has just a few simple steps.

- ▶ Calculate the difference for each pair of observations
- ▶ Rank the absolute values of these differences from smallest to largest (drop 0's; assign an average rank to ties); let n be the number of non-zero differences
- ▶ Assign each rank a $+$ or $-$ depending on the original sign of the difference
- ▶ Add up all positive ranks; add up all negative ranks; let T be equal to the smaller sum
- ▶ Calculate the z-score $z_T = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$ and obtain p-value from the normal distribution. If $n \leq 12$ use Table A.6 from P&G instead.

Example: Middle Ear Effusion

A common symptom of otitis media in young children is the prolonged presence of fluid in the middle ear, known as *middle-ear effusion*. The presence of fluid may result in temporary hearing loss and interfere with normal learning skills in the first 2 years of life. One hypothesis is that babies who are breastfed for at least one month build some immunity and have less prolonged effusion than their bottle-fed counterparts. A small study of 14 babies is set up, with babies matched one-to-one by age, gender, socioeconomic status, and health condition. One member of the pair is a breastfed baby, and the other member is a bottle-fed baby. The outcome is the duration (days) of middle-ear effusion after the first episode of otitis media.

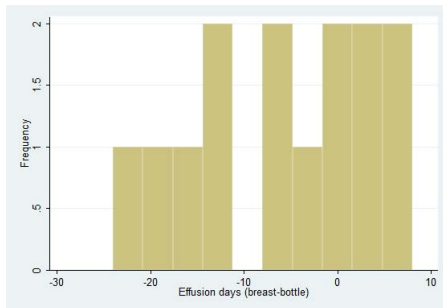
Example: Middle Ear Effusion

Table 1: Duration of effusion study

Pair Number	Duration of Effusion for Each Pair (Days)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Breastfed	26	3	12	28	7	39	12	30	7	15	65	10	19	11
Bottle-fed	18	7	6	33	7	57	29	28	8	27	78	17	16	35

What is H_0 ?

Example: Middle Ear Effusion



Example: Middle Ear Effusion

Table 2: Duration of effusion study

Pair Number	Duration of Effusion for Each Pair (Days)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Breastfed	26	3	12	28	7	39	12	30	7	15	65	10	19	11
Bottle-fed	18	7	6	33	7	57	29	28	8	27	78	17	16	35
Difference	8	-4	6	-5	0	-18	-17	2	-1	-12	-13	-7	3	-24
Difference	8	4	6	5	0	18	17	2	1	12	13	7	3	24
Rank	8	4	6	5	NA	12	11	2	1	9	10	7	3	13
Signed Rank	8	-4	6	-5	NA	-12	-11	2	-1	-9	-10	-7	3	-13

$n = 13$ nonzero ranks

Sum of positive ranks: $8+6+2+3=19$

$T = 19$ and $z_T = \frac{19 - \frac{13(14)}{4}}{\sqrt{\frac{13(14)(27)}{24}}} = \frac{19 - 45.5}{14.309} = -1.85, p > 0.05$. What do you conclude?

Wilcoxon Signed Rank Test in Stata

```
. signrank breast=bottle
```

Wilcoxon signed-rank test

sign	obs	sum ranks	expected
positive	4	23	52
negative	9	81	52
zero	1	1	1
all	14	105	105

unadjusted variance 253.75

adjustment for ties 0.00

adjustment for zeros -0.25

adjusted variance 253.50

Ho: breastfed = bottle

z = -1.821

Prob > |z| = 0.0685

Note: Stata uses a slightly different method, which is why the results are a little different. Basically, it ranks the zero but then makes an adjustment for the zeros.

Wilcoxon Rank Sum Test

- ▶ Nonparametric counterpart of the two-sample t-test
- ▶ Assumes two samples are independent
- ▶ Does not require normality or equal variance of groups
- ▶ Assumes two distributions have roughly the same shape
- ▶ Evaluates H_0 : the median difference between groups is 0
- ▶ Also called Mann-Whitney U test or Wilcoxon-Mann-Whitney test

Wilcoxon Rank Sum Test

The Wilcoxon Rank Sum Test is an extension of the Wilcoxon signed rank test for independent samples. Instead of taking differences, the observations are lumped together and ranked, with then the sum of the ranks calculated by group. The test statistic is the smaller sum of the ranks. A z-score is constructed for moderate-to-large samples as described in P&G (more than 10 in each group); for small samples exact tabled values may be used (see table under Resources for this lecture).

Example: Exercise Capacity and Coronary Artery Disease

We consider data from a two-group study of exercise capacity. Two groups of men, one with diagnosed three-vessel coronary artery disease (3VD), and the other group of men with suspected disease (SD) in one or more vessels. The total time (in seconds) men were able to exercise on a treadmill, set to increase in speed and slope according to a set schedule, is below.

3VD times: 864, 636, 638, 708, 786, 600, 1320, 750, 594, 750

SD times: 1014, 684, 810, 990, 840, 978, 1002, 1110

3VD median: 729

SD median: 984

Example: Exercise Capacity and Coronary Artery Disease

Value	Rank	Group	Value	Rank	Group
594	1	3VD	810	10	SD
600	2	3VD	840	11	SD
636	3	3VD	864	12	3VD
638	4	3VD	978	13	SD
684	5	SD	990	14	SD
708	6	3VD	1002	15	SD
750	7.5	3VD	1014	16	SD
750	7.5	3VD	1110	17	SD
786	9	3VD	1320	18	3VD

Sum of ranks in SD group ($n=8$): 101

Sum of ranks in 3VD group ($n=10$): 70

Appendix : Wilcoxon Rank-Sum Table

Probabilities relate to the distribution of W_A , the rank sum for group A when $H_0 : A = B$ is true. The tabulated value for the **lower tail** is the largest value of w_A for which $\text{pr}(W_A \leq w_A) \leq \text{prob}$. The tabulated value for the **upper tail** is the smallest value of w_A for which $\text{pr}(W_A \geq w_A) \leq \text{prob}$.

		Lower Tail						Upper Tail					
n_A	n_B	<i>prob</i>						<i>prob</i>					
		.005	.01	.025	.05	.10	.20	.20	.10	.05	.025	.01	.005
4	4			10	11	13	14	22	23	25	26		
	5		10	11	12	14	15	25	26	28	29	30	
	6		10	11	12	13	15	27	29	31	32	33	34
	7		10	11	13	14	16	30	32	34	35	37	38
	8		11	12	14	15	17	32	35	37	38	40	41
	9		11	13	14	16	19	35	37	40	42	43	45
	10		12	13	15	17	20	37	40	43	45	47	48
	11		12	14	16	18	21	40	43	46	48	50	52
	12		13	15	17	19	22	42	46	49	51	53	55
5	5	15	16	17	19	20	22	33	35	36	38	39	40
	6	16	17	18	20	22	24	36	38	40	42	43	44
	7	16	18	20	21	23	26	39	42	44	45	47	49
	8	17	19	21	23	25	28	42	45	47	49	51	53
	9	18	20	22	24	27	30	45	48	51	53	55	57
	10	19	21	23	26	28	32	48	52	54	57	59	61
	11	20	22	24	27	30	34	51	55	58	61	63	65
	12	21	23	26	28	32	36	54	58	62	64	67	69
6	6	23	24	26	28	30	33	45	48	50	52	54	55
	7	24	25	27	29	32	35	49	52	55	57	59	60
	8	25	27	29	31	34	37	53	56	59	61	63	65
	9	26	28	31	33	36	40	56	60	63	65	68	70
	10	27	29	32	35	38	42	60	64	67	70	73	75
	11	28	30	34	37	40	44	64	68	71	74	78	80
	12	30	32	35	38	42	47	67	72	76	79	82	84
7	7	32	34	36	39	41	45	60	64	66	69	71	73
	8	34	35	38	41	44	48	64	68	71	74	77	78
	9	35	37	40	43	46	50	69	73	76	79	82	84
	10	37	39	42	45	49	53	73	77	81	84	87	89
	11	38	40	44	47	51	56	77	82	86	89	93	95
	12	40	42	46	49	54	59	81	86	91	94	98	100
8	8	43	45	49	51	55	59	77	81	85	87	91	93
	9	45	47	51	54	58	62	82	86	90	93	97	99
	10	47	49	53	56	60	65	87	92	96	99	103	105
	11	49	51	55	59	63	69	91	97	101	105	109	111
	12	51	53	58	62	66	72	96	102	106	110	115	117
9	9	56	59	62	66	70	75	96	101	105	109	112	115
	10	58	61	65	69	73	78	102	107	111	115	119	122
	11	61	63	68	72	76	82	107	113	117	121	126	128
	12	63	66	71	75	80	86	112	118	123	127	132	135
10	10	71	74	78	82	87	93	117	123	128	132	136	139
	11	73	77	81	86	91	97	123	129	134	139	143	147
	12	76	79	84	89	94	101	129	136	141	146	151	154
11	11	87	91	96	100	106	112	141	147	153	157	162	166
	12	90	94	99	104	110	117	147	154	160	165	170	174
12	12	105	109	115	120	127	134	166	173	180	185	191	195

Example: Exercise Capacity and Coronary Artery Disease

The sum of the ranks in the SD group ($n=8$) was 101, and the 3VD group had 10 subjects with sum of ranks 70. The table indicates that we would reject H_0 for a sum of ranks in the smaller group at or below 53 or at 99 or higher. We reject H_0 and conclude the two groups do not have the same median. The median in the SD group is higher.

Note: if we use the normal approximation (not valid in this case as one group has fewer than 10 subjects) in Stata, we reject H_0 .

Example: China Data

When we learned about the two-sample t-test, we tested whether the average BMI among smokers was the same as that among nonsmokers. We can now evaluate whether the median BMI is the same for smokers (estimated median 22.8) and nonsmokers (22.9), using a test that does not require normality of BMI.

```
. ranksum bmi, by(smokeyn)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

smokeyn	obs	rank sum	expected
0	6460	28597549	28372320
1	2323	9977387.5	10202616
combined	8783	38574936	38574936


```
unadjusted variance    1.098e+10
adjustment for ties    -1152.4545
-----
adjusted variance      1.098e+10

Ho: bmi(smokeyn==0) = bmi(smokeyn==1)
      z =      2.149
Prob > |z| =      0.0316
```

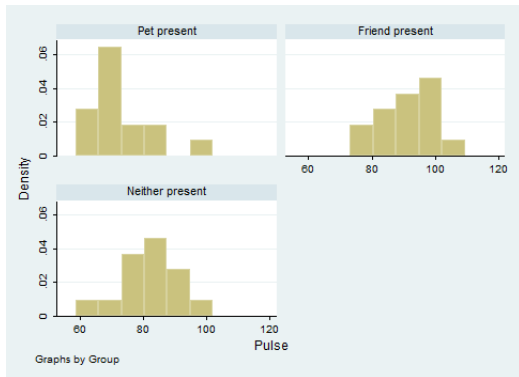
What do we conclude?

Kruskal-Wallis Test

The Kruskal-Wallis Test is the nonparametric version of ANOVA, generalizing the Wilcoxon rank sum test to more than 2 groups.

Kruskal-Wallis Test for Pet Data

Consider the study of pulse rate from the ANOVA lecture. Pulse rates are not necessarily normally distributed within groups.



Kruskal-Wallis Test for Pet Data

```
. histogram Pulse, by(Group)
```

```
. kwallis Pulse, by(Group)
```

Kruskal-Wallis equality-of-populations rank te

Group	Obs	Rank Sum
Pet pres	15	190.00
Friend p	15	495.00
Neither	15	350.00

```
chi-squared =      17.990 with 2 d.f.  
probability =      0.0001
```

```
chi-squared with ties =      17.990 with 2 d.f.  
probability =      0.0001
```

What do you conclude?

Pairwise Comparisons

We can look at pairwise differences using the Wilcoxon Rank-Sum test. Note it does not automatically adjust for multiple comparisons, so we'd need to do that on our own if desired. First we need to generate additional variables with *only* two levels.

```
. generate petfriend=.  
(45 missing values generated)  
  
. generate petneither=.  
(45 missing values generated)  
  
. generate friendneither=.  
(45 missing values generated)  
  
. replace petfriend=1 if Group==1  
(15 real changes made)  
  
. replace petfriend=2 if Group==2  
(15 real changes made)  
  
. replace petneither=1 if Group==1  
(15 real changes made)  
  
. replace petneither=3 if Group==3  
(15 real changes made)  
  
. replace friendneither=2 if Group==2  
(15 real changes made)  
  
. replace friendneither=3 if Group==3  
(15 real changes made)
```

Pairwise Comparisons: Pet versus Friend

```
. ranksum Pulse, by(petfriend)
```

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

petfriend	obs	rank sum	expected
1	15	139	232.5
2	15	326	232.5
combined	30	465	465

unadjusted variance **581.25**

adjustment for ties **0.00**

adjusted variance **581.25**

Ho: Pulse(petfri~d==1) = Pulse(petfri~d==2)

z = **-3.878**

Prob > |z| = **0.0001**

Pairwise Comparisons: Pet versus Neither

```
. ranksum Pulse, by(petneither)
```

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

petneither	obs	rank sum	expected
1	15	171	232.5
3	15	294	232.5
combined	30	465	465

unadjusted variance **581.25**

adjustment for ties **0.00**

adjusted variance **581.25**

Ho: Pulse(petnei~r==1) = Pulse(petnei~r==3)

 z = **-2.551**

 Prob > |z| = **0.0107**

Pairwise Comparisons: Friend versus Neither

```
. ranksum Pulse, by(friendneither)
```

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

friendneit~r	obs	rank sum	expected
2	15	289	232.5
3	15	176	232.5
combined	30	465	465

unadjusted variance **581.25**

adjustment for ties **0.00**

adjusted variance **581.25**

Ho: Pulse(friend~r==2) = Pulse(friend~r==3)

z = **2.344**

Prob > |z| = **0.0191**

Pairwise Comparisons

Which groups are different? Does doing a Bonferroni correction affect your answer?

Why Not Always Go Nonparametric?



Nonparametric methods are desirable because they do not require as many restrictive assumptions as parametric ones. However, this flexibility comes at a price – *if* the assumptions underlying a parametric test are satisfied, the nonparametric test is less powerful than the comparable parametric technique. For example, if the data are normal, the power of the Wilcoxon signed-rank test is $\approx 95\%$ of that of the t-test.