

**22S:172****SAS for Data Management, Analysis,  
and Reporting**

Lecture 3

Introduction to Nonparametric Methods  
 Analysis of Variance and Wilcoxon Rank-Sum Test  
 June 23, 2004

Kate Cowles  
 374 SH  
 kcowles@stat.uiowa.edu

**“Nonparametric” or “distribution-free”  
statistical methods**

- allow for testing hypotheses that are not statements about population parameter values
- may be used when the form of the distribution of the sampled population is unknown
- can be used when data being analyzed consist merely of rankings or classifications
  - i.e. when arithmetic operations required for parametric procedures cannot be done
  - example: data on patient conditions reported as “better,” “same,” or “worse”

**Parametric methods**

- based on the assumption that the population(s) from which our samples are drawn follow a distribution, the general form of which is known
  - e.g. normal or binomial
- research interest is in estimating, or testing a hypothesis about, one or more population parameters
- examples: z tests, t tests, and ANOVA for making inference about means of populations assumed to be normal

**Example for the Sign Test**

- We wish to compare the effectiveness of two ointments (A, B) in reducing sunburn in people whose skin is sensitive to sunlight.
- For each person in the study, we randomly select either the left arm or the right arm and apply ointment A. We then apply ointment B to the same area of the other arm.
- We then expose the person to 1 hour of sunlight and compare the two arms with respect to degree of redness.
- We can make only the following qualitative assessments:
  1. “A” arm is not as red as “B” arm.
  2. “A” arm is redder than “B” arm.
  3. Arms are equally red.

How might we compare the effectiveness of the two ointments *if we were able to measure redness on a quantitative scale?*

In the situation described here, we cannot observe the actual values of within-person differences in redness between the A arm and the B arm.

What we can observe are the *signs* of the differences:

1. "A" arm is not as red as "B" arm (+)
2. "A" arm is redder than "B" arm (-)
3. Arms are equally red (0)

To carry out the sign test:

- Ignore the pairs (or observations) with difference of 0.
- Denote the number of remaining pairs as  $n$ .
- Count the number of plus signs, and denote it  $D$ .
- Note that under the null hypothesis, we would expect approximately equal numbers of plus and minus signs.
  - more precisely, under the null hypothesis,  $D$  follows a binomial distribution with success probability  $p = 1/2$  and number of trials  $n$
  - This binomial distribution has

$$mean = np = \frac{n}{2}$$

$$standarddeviation = \sqrt{np(1-p)} = \sqrt{\frac{n}{4}}$$

## The Sign Test

The null hypothesis of the sign test is that in the underlying population of differences, the median difference  $M$  is 0.

$$H_0 : M = 0.$$

The alternative hypothesis may be either one-sided or two-sided.

$$H_0 : M > 0$$

$$H_0 : M < 0$$

$$H_0 : M \neq 0$$

- We must evaluate how likely we would have been to obtain a value of  $D$  as extreme as what we got, or more extreme, if the null is true.
- Your textbook gives the test statistic for use with a normal approximation to the binomial distribution. This is appropriate for use if  $n \geq 20$ . The value is compared to the standard normal distribution.
- Otherwise, we will use the binomial distribution directly.

## The sign test for the skin ointment data

We wish to do a two-sided test, i.e.

$$H_a : M \neq 0$$

at the  $\alpha = .05$  significance level.

The results for 45 subjects are:

1. 22 people had the “A” arm less red (+)
2. 18 people had the “B” arm less red (-)
3. 5 people had no difference (0)

- $n = 45 - 5 = 40$
- $D = 22$
- normal approximation is valid because  $n \geq 20$ .

$$z_+ = \frac{D - (n/2)}{\sqrt{n/4}}$$

So again, we cannot reject  $H_0$ . We conclude that the data do not provide evidence that one ointment is better than the other.

$$\begin{aligned} &= \frac{22 - 20}{\sqrt{10}} \\ &= 0.632 \end{aligned}$$

For a 2-sided test, we must compare this value to the .025 cutoff for the standard normal distribution, which is 1.96.

Because  $0.632 < 1.96$ , we cannot reject  $H_0$ .

Equivalently, we can determine the p-value of our test by finding  $P(z > 0.632) \approx .264$ .

- This would be the p-value for a 1-sided test.
- To find the p-value for our 2-sided test, we multiply by 2.

$$p = 2(.264) = .528 > \alpha = .05$$

## The sign test with small sample size

Suppose that instead of 40 patients with non-zero differences, we had had

1. 5 people had the “A” arm less red (+)
2. 3 people had the “B” arm less red (-)
3. 37 people had no difference (0)

Then

- $n = 45 - 37 = 8$
- $D = 5$
- normal approximation is inappropriate because  $n < 20$ .
  - we will do exact calculation of the p-value using the binomial distribution

Because  $D > n/2 = 4$ , we will compute

$$\begin{aligned} P(D \geq 5|H_0) &= P(D = 5) + P(D = 6) \\ &\quad + P(D = 7) + P(D = 8) \\ &= .2188 + .1094 + .0313 + .0039 \\ &= 0.3634 \end{aligned}$$

This is a one-sided p-value. We must multiply by 2 to get the approximate 2-sided p-value.

$$2(0.3634) = 0.7268 > .05$$

So again we would not reject  $H_0$ .

## The Wilcoxon Signed-Rank Test

- for single sample or paired samples
- useful when the population distribution is not normal and the sample size is not large
  - of the within-pair differences in paired sample case or of individual values in single sample case
- makes use of the magnitudes of the differences as well as their signs

## More on the sign test

- Can be used with single-sample or paired-sample problems
- Frees us from having to make any assumptions about the underlying distribution of differences
- If we have any information about the magnitude of the individual differences, the sign test wastes it.

## Example

Kashima, Baker, and Landen (1988) studied whether media-based instruction could help the parents of mentally handicapped children become more effective at teaching their children self-help skills.

As part of the study, 17 families participated in a training program. Before and after the training program, the primary parent took the Behavioral Vignettes test, which assesses knowledge of behavioral modification principles. A higher score indicates greater knowledge.

The following are the pre- and post-test training scores for 12 of their families:

Pre	Post
7	11
6	14
10	16
16	17
8	9
13	15
8	9
14	17
16	20
11	12
12	14
13	15

May we conclude from these data that the training program increases knowledge of behavior modification principles? (We will test at the  $\alpha = .01$  level.)

### Steps in the Wilcoxon signed-rank procedure

1. Select a random sample of  $n$  pairs of observations.
2. Compute the difference  $d_i$  in each pair of observations. Delete all pairs in which  $d_i = 0$ , and reduce  $n$  accordingly.
3. Ignoring the signs of the  $d_i$ s, rank their absolute values from smallest to largest. When there are ties in absolute values, assign each tied value the mean of the rank positions the tied values occupy.
4. Assign to each rank the sign of the  $d_i$  that yields that rank.

### Hypotheses of the Wilcoxon Signed Rank Test

The null hypothesis is that, in the underlying population of differences among pairs, the median difference is equal to 0.

$$H_0 : M_d = 0$$

The alternative hypothesis may be one- or two-sided.

$$H_a : M_d > 0$$

$$H_a : M_d < 0$$

$$H_a : M_d \neq 0$$

If we define our differences as post - pre, then our alternative would be:

$$H_a : M_d > 0$$

5. Find  $T_+$ , the sum of the ranks with positive signs, and  $T_-$ , the sum of the ranks with negative signs.
6. Let the test statistic  $T$  equal the smaller of  $T_+$  and  $T_-$ .

## Pre- and Post-Test example

Pre	Post	d_i	Rank
7	11	4	9.5
6	14	8	12
10	16	6	11
17	16	-1	-2.5
8	9	1	2.5
13	15	2	6
8	9	1	2.5
14	17	3	8
16	20	4	9.5
11	12	1	2.5
12	14	2	6
13	15	2	6

The sum of the negative ranks is  $T = 2.5$ .

```

data whatever ;
input pre post ;
diff = post - pre ;
datalines ;
 7 11
 6 14
10 16
16 17
 8  9
13 15
 8  9
14 17
16 20
11 12
12 14
13 15
;
run ;

```

## SAS for the Wilcoxon Signed Rank Test

- carried out automatically by `proc univariate`
- SAS computes a slightly different form of the test statistic

$$S = \Sigma(\text{positive ranks}) - \frac{n(n+1)}{4}$$

recalling that  $n$  is the number of differences whose value is not equal to 0.

- computes p-value in two different ways depending on sample size
  - if  $n \leq 20$ , p-value is computed from each distribution of  $S$ , which can be enumerated under null hypothesis that distribution is symmetric around 0
  - when  $n > 20$  approximate  $S$  is compared to approximate t distribution

```

proc univariate ;
var diff ;
run ;

```

The UNIVARIATE Procedure  
Variable: diff

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value-----
Student's t	t 4.521908	Pr >  t  0.0009
Sign	M 6	Pr >=  M  0.0005
Signed Rank	S 39	Pr >=  S  0.0005

Note: For a 1-sided p-value, we would divide the 2-sided p-value by 2.

## Interpreting the results

- Recall that we wanted to determine whether the audiovisual instruction improved parent's test scores.
- The null and alternative hypotheses regarding the median difference (that is, the median of post - pre), are

$$H_0 : M_d = 0$$

$$H_a : M_d > 0$$

- Can we reject  $H_0$  at the .01 significance level?
- What does this mean with respect to the research question?

## The Wilcoxon Rank Sum Test

- used to compare nonparametrically two samples that have been drawn from independent populations
  - nonparametric analog of two-independent-sample t-test
- also called Mann-Whitney test, Mann-Whitney U test, and Mann-Whitney-Wilcoxon test

## Sign Test in SAS

- Note that `proc univariate` also automatically carries out the sign test
- its version of sign test statistic is

$$M - \frac{n^+ - n^-}{2}$$

- use sign test if sample size is small and it is unreasonable to assume that population distribution is *symmetric*
- sign test p-value will often be a little larger than that of the Wilcoxon signed rank test (not so in this case)

## Assumptions of the Wilcoxon Rank Sum Test

- Two samples, of sizes  $n$  and  $m$ , have been drawn independently and randomly from their respective populations
- The measurement scale is at least ordinal
- The variable of interest is continuous
- If the populations differ, they differ only with respect to their medians
  - i.e., otherwise their shapes are approximately the same

## Example: a question in pharmacokinetics

- Is total plasma clearance of cefpiramide different in healthy people vs. patients with alcoholic cirrhosis? Demotes-Mainard et al. (1991) measured total plasma clearance (ml/min)
- following a single 1-gram intravenous injection of cefpiramide in 10 healthy volunteers and 10 patients with alcoholic cirrhosis.
- They chose to use a non-parameteric method of analysis because their medical knowledge indicated that plasma clearance of drugs tends to have a skewed (not normal!) distribution

Case number	CIRR	CLEAR
1	0.000	21.700
2	0.000	29.300
3	0.000	25.300
4	0.000	22.800
5	0.000	21.300
6	0.000	31.200
7	0.000	29.200
8	0.000	28.700
9	0.000	17.200
10	0.000	25.700
11	1.000	14.600
12	1.000	18.100
13	1.000	12.300
14	1.000	8.800
15	1.000	10.300
16	1.000	8.500
17	1.000	29.300
18	1.000	8.100
19	1.000	6.900
20	1.000	7.900

Can we conclude at the  $\alpha = .01$  significance level that median clearance rate is different in healthy patients vs. those with alcoholic cirrhosis?

## Hypotheses for the Wilcoxon Rank Sum Test

$$H_0 : M_1 = M_2$$

The alternative hypothesis may be one- or two-sided.

$$H_a : M_1 > M_2$$

$$H_a : M_1 < M_2$$

$$H_a : M_1 \neq M_2$$

## Procedure for the Wilcoxon Rank Sum Test

- Combine the two samples into one large group, and sort values from smallest to largest.
- Rank the values. When there are ties in absolute values, assign each tied value the mean of the rank positions the tied values occupy.
- Sum the ranks *within each original sample*
- The test statistic is  $W$ , the smaller of the two sums.

## Ranked values for clearance example

```

Case
number CIRR CLEAR
1 0.000 21.700 12.0
2 0.000 29.300 18.5
3 0.000 25.300 14.0
4 0.000 22.800 13.0
5 0.000 21.300 11.0
6 0.000 31.200 20.0
7 0.000 29.200 17.0
8 0.000 28.700 16.0
9 0.000 17.200 9.0
10 0.000 25.700 15.0
11 1.000 14.600 8.0
12 1.000 18.100 10.0
13 1.000 12.300 7.0
14 1.000 8.800 5.0
15 1.000 10.300 6.0
16 1.000 8.500 4.0
17 1.000 29.300 18.0
18 1.000 8.100 3.0
19 1.000 6.900 1.0
20 1.000 7.900 2.0

```

```

proc npar1way wilcoxon ;
class cirr ;
var clear ;
run ;

```

## The NPARIWAY Procedure

Wilcoxon Scores (Rank Sums) for Variable clear  
Classified by Variable cirr

cirr	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	10	145.50	105.0	13.223782	14.550
1	10	64.50	105.0	13.223782	6.450

Average scores were used for ties.

## Wilcoxon Two-Sample Test

Statistic	145.5000
Normal Approximation Z	3.0249
One-Sided Pr > Z	0.0012
Two-Sided Pr >  Z	0.0025
t Approximation	
One-Sided Pr > Z	0.0035
Two-Sided Pr >  Z	0.0070

Z includes a continuity correction of 0.5.

## The Wilcoxon Rank Sum test in SAS

- use proc npar1way

```

data clear ;
input id cirr clear ;
datalines ;
1 0.000 21.700
2 0.000 29.300
3 0.000 25.300
4 0.000 22.800
5 0.000 21.300
6 0.000 31.200
7 0.000 29.200
8 0.000 28.700
9 0.000 17.200
10 0.000 25.700
11 1.000 14.600
12 1.000 18.100
13 1.000 12.300
14 1.000 8.800
15 1.000 10.300
16 1.000 8.500
17 1.000 29.300
18 1.000 8.100
19 1.000 6.900
20 1.000 7.900
;
run ;

```

## Comparing more than two population means

Example: Does the presence of pets or friends affect responses to stress?

- Allen, Blascovich, Tomaka, and Kelsey, 1988, *Journal of Personality and Social Psychology*
- subjects: 45 women who described themselves as dog lovers
- randomly assigned to three groups: to do a stressful task
  1. alone
  2. with a good friend present
  3. with their dog present
- Subjects' mean heart rate during the task was one measure of the effect of stress.

## Goal: to compare population means under three different “treatments”

- a *three*-independent-sample problem
- Call the population mean heart rates  $\mu_1$  for when pets are present,  $\mu_2$  for when friends are present, and  $\mu_3$  for when women perform task alone: then

$$- H_0 : \mu_1 = \mu_2 = \mu_3$$

$$- H_a : \mu_1 \neq \mu_2 \text{ or } \mu_1 \neq \mu_3 \text{ or } \mu_2 \neq \mu_3$$

\* not one-sided or 2-sided

SAS descriptive statistics:

Analysis Variable : BEATS

group	N		Mean	Std Dev	Minimum	Maximum
	Obs	N				
C	15	15	82.5240667	9.2415747	62.6460000	99.0460000
F	15	15	91.3251333	8.3411341	76.9080000	102.1540000
P	15	15	73.4830667	9.9698202	58.6920000	97.5380000

## Multiple comparisons procedures in statistics

- issue: how to do many comparisons at once with some overall measure of confidence in all our conclusions
- two steps
  - overall test of whether there is good evidence of *any* differences among parameters we wish to compare
  - follow-up analysis to decide which of parameters differ and to estimate size of differences

## Step one: One-Way Analysis of Variance (ANOVA)

- step one (overall test) for *some* difference among 3 or more population means
- uses an *F test* to compute a p-value

## Main idea of ANOVA

What matters is how far apart sample means are *relative to variability of individual observations*.

- F statistic

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

- compare to a cutoff value in an **F distribution**

Notation:

- $I$  = number of different populations whose means we are studying
- $n_i$  = number of observations in sample from  $i$ th population
- $N$  = total number of observations in all samples combined

## Assumptions for One-Way ANOVA

- We have  $I$  independent simple random samples, one from each of  $I$  populations.
- Each population  $i$  has a normal distribution with unknown mean  $\mu_i$ .
  - As with  $t$ -tests, if sample sizes are large enough in each sample, Central Limit Theorem says inference based on sample means is OK even if population distributions are not exactly normal.

## F distributions

- many different F distributions, identified by two parameters
  - numerator degrees of freedom =  $I - 1$
  - denominator degrees of freedom =  $N - I$

- All of the populations have the same standard deviation  $\sigma$  (unknown)
  - unlike  $t$ -tests, there is no general procedure when population standard deviations are not assumed to be equal
  - rough rule of thumb: if largest sample standard deviation is no more than twice the smallest sample standard deviation, then population standard deviations probably are close enough to equal that ANOVA procedure is OK

## We can check assumptions using proc means and proc univariate

```
options linesize = 79 ;

data pet ;
infile '/temp/pet.dat' ;
input group $ beats ;
run ;

proc sort data = pet ;
by group ;
run ;

proc means data = pet ;
by group ;
var beats ;
run ;

proc univariate normal plot data = pet ;
by group ;
var beats ;
```

The MEANS Procedure

Analysis Variable : beats

group	N		Mean	Std Dev	Minimum	Maximum
	Obs	N				
C	15	15	82.5240667	9.2415747	62.6460000	99.0460000
F	15	15	91.3251333	8.3411341	76.9080000	102.1540000
P	15	15	73.4830667	9.9698202	58.6920000	97.5380000

----- group=C -----

The UNIVARIATE Procedure

Variable: beats

Moments

	N	15	Sum Weights	15
Mean	82.5240667	82.5240667	Sum Observations	1237.861
Std Deviation	9.24157468	9.24157468	Variance	85.4067026
Skewness	-0.4775032	-0.4775032	Kurtosis	0.32351507
Uncorrected SS	103349.018	103349.018	Corrected SS	1195.69384
Coeff Variation	11.1986419	11.1986419	Std Error Mean	2.38616432

Basic Statistical Measures

	Location	Variability
--	----------	-------------

Mean	82.52407	Std Deviation	9.24157
Median	84.73800	Variance	85.40670
Mode	.	Range	36.40000

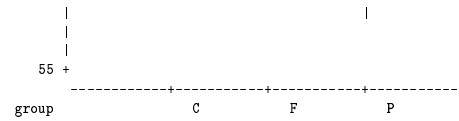
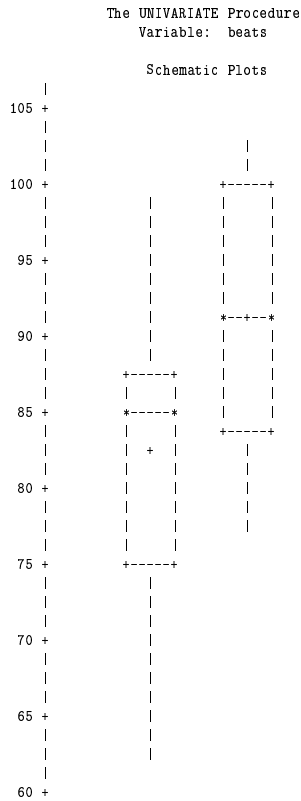
Tests for Normality

Test	--Statistic--	----p Value-----
Shapiro-Wilk	W 0.970236	Pr < W 0.8615
Kolmogorov-Smirnov	D 0.185622	Pr > D >0.1500
Cramer-von Mises	W-Sq 0.052736	Pr > W-Sq >0.2500
Anderson-Darling	A-Sq 0.285759	Pr > A-Sq >0.2500

Stem Leaf	#	Boxplot
9 9	1	
9 02	2	
8 555778	6	+-----+
8 0	1	+
7 58	2	+-----+
7 13	2	
6		
6 3	1	
-----+-----+-----+		

Multiply Stem.Leaf by 10\*\*\*1

Proc univariate provides side-by-side boxplots at the end of the output when a **by** statement is used. (It does not do this when a **class** statement is used to get separate output for different groups.)



## So go ahead and do ANOVA

Dogs, friends, and stress example:

```
proc anova data = pet ;
class group ;
model beats = group ;
run ;
```

Analysis of Variance Procedure

Class	Levels	Values
GROUP	3	C F P

Number of observations in data set = 45

Analysis of Variance Procedure

Dependent Variable: BEATS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2387.6889920	1193.8444960	14.08	0.0001
Error	42	3561.2994916	84.7928450		
Corrected Total	44	5948.9884836			

	R-Square	C.V.	Root MSE	BEATS Mean
	0.401360	11.16915	9.2083030	82.444089

Source	DF	Anova SS	Mean Square	F Value	Pr > F
GROUP	2	2387.6889920	1193.8444960	14.08	0.0001

## Step two: individual t-tests with correction for multiple comparisons

This is the *follow-up* test.

- should be carried out *only* if the F test from one-way ANOVA is significant at the chosen significance level.

Goal: to set the *overall* probability of committing a type I error at  $\alpha$  when doing pairwise comparisons of  $k$  different means

- we will perform  $\binom{k}{2}$  two-independent-sample t-tests
- we will conduct each one at the significance level

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}$$

- This is called the *Bonferroni correction*
  - very conservative

Dogs, friends, and stress example

- There are  $k = 3$  samples, so there are  $\binom{k}{2} = 3$  different pairs to compare.
- To get an overall significance level  $\alpha = .05$  on all 3 tests considered together, we conduct each one at

$$\alpha^* = \frac{.05}{3} = .0167$$

- That is, we would consider the difference between two population means to be significantly different from zero at the .05 level only if the p-value for the the t-test for that pair was less than .0167.

- Equivalently, we could multiply the p-value from each t-test by 3.
  - \* If the result was less than .05, we would consider the difference between two population means to be significantly different from zero at the .05 level

SAS does the adjusting and prints a grouped list of the classes. Means with the same letter are not significantly different at the specified alpha level.

```
proc anova data = pet ;
class group ;
model beats = group ;
means group / bon alpha = .05 ;
run ;
```

Analysis of Variance Procedure

Bonferroni (Dunn) T tests for variable: BEATS

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 42 MSE= 84.79285  
Critical Value of T= 2.49  
Minimum Significant Difference= 8.3847

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	GROUP
A	91.325	15	F
B	82.524	15	C
C	73.483	15	P

## The Kruskal-Wallis test

- ANOVA extends idea of 2-independent-sample t-test to more than 2 independent samples
- Wilcoxon Rank-Sum test is non-parametric way of comparing centers of population distributions using two independent samples
- Kruskal-Wallis test extends idea of Wilcoxon Rank-Sum test to more than 2 independent samples
- like Wilcoxon Rank-Sum, Kruskal-Wallis has less *power* than ANOVA, but it does not depend on such strong distributional assumptions
- SAS `proc npar1way` is used for the Kruskal-Wallis test

## Kruskal Wallis test for pets and stress data

Note: the Kruskal Wallis test is not the best choice for this dataset, because both subject-matter knowledge and inspection of sample data suggest that assumption of population normality probably is reasonable.

```
proc npar1way wilcoxon ;
class group ;
var beats ;
run ;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable beats  
Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
C	15	350.0	345.0	41.533119	23.333333
F	15	495.0	345.0	41.533119	33.000000
P	15	190.0	345.0	41.533119	12.666667

Kruskal-Wallis Test

Chi-Square	17.9903
DF	2
Pr > Chi-Square	0.0001