AMS 572 Data Analysis I
Nonparametric Statistical Methods

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Nonparametric statistical methods

- Nonparametric methods are distribution free statistical methods.
- So far the methods that we introduced in earlier chapters works under normal distribution assumption or large sample theory (CLT)
- When the sample size is small and the data is not normally distributed, the proposed t-tests is not applicable.

Nonparametric inference for single samples

Sign Test

- Suppose $Y_1, \ldots, Y_n$ iid continuous F with median $\tilde{\mu}$
- The median is used as it is a better measure of the center than the mean for nonnormal distributions
- Hypotheses

$$H_0 : \tilde{\mu} = \tilde{\mu}_0 \text{ (median=some value)}$$

$$H_a : \tilde{\mu} > \tilde{\mu}_0$$

- If the true median is $\tilde{\mu}_0$ and F continuous,

$$\Pr[Y < \tilde{\mu}_0] = \Pr[Y > \tilde{\mu}_0] = 0.5$$

for a randomly selected observation Y

Sign Test

- Let S_+ be the number of obs $> \tilde{\mu}_0$
- Similarly, S_- be the number of obs $< \tilde{\mu}_0$. Note that $S_- = n - S_+$.

- Reject H_0 if S_+ is large or equivalently S_- is small
- Under H_0

$$S_+ \sim Bin(n, 0.5)$$

$$\Pr[S_+ \leq s_+] = \sum_{i=0}^{s_+} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \frac{1}{2^n} \sum_{i=0}^{s_+} \binom{n}{i}$$

Sign Test

- Hypotheses

$$H_0 : \tilde{\mu} = \tilde{\mu}_0$$

$$H_a : \tilde{\mu} > \tilde{\mu}_0$$

- Equivalently

$$H_0 : p = 0.5$$

$$H_a : p > 0.5$$

Sign Test

- Critical region

$$C_\alpha = \{s_+ : s_+ \geq b_\alpha\}$$

where

$$\Pr[S_+ \geq b_\alpha | H_0] \leq \alpha$$

Sign Test

- P-value $= P(S_+ \geq s_+) = \frac{1}{2^n} \sum_{i=s_+}^{n} \binom{n}{i}$
- Reject H_0 if p-value $\leq \alpha$.

Sign Test

- One tailed: $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs $H_a : \tilde{\mu} > \tilde{\mu}_0$
  P-value $= P(S_+ \geq s_+) = \frac{1}{2^n} \sum_{i=s_+}^{n} \binom{n}{i}$
- One tailed: $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs $H_a : \tilde{\mu} < \tilde{\mu}_0$
  P-value $= P(S_- \geq s_-) = \frac{1}{2^n} \sum_{i=s_-}^{n} \binom{n}{i}$
- Two tailed: $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs $H_a : \tilde{\mu} \neq \tilde{\mu}_0$
  P-value $= 2 \sum_{i=s_{\max}}^{n} \frac{1}{2^n} \binom{n}{i}$ where $s_{\max} = \max(s_+, s_-)$

Example: Consider the following thermostat data
202.2, 203.4, 200.5, 202.5, 206.3, 198.0, 203.7, 200.8, 201.3, 199.0
Test if the median setting is different from the design setting of 200
This is example 14.1 of your textbook

```
> 2*sum(dbinom(8:10,10,0.5))
[1] 0.109375
>
> ## Alternative way ###
> x <- c(202.2, 203.4, 200.5, 202.5, 206.3, 198.0,
203.7, 200.8, 201.3, 199.0)
> library("BSDA")
> SIGN.test(x,md=200)

One-sample Sign-Test

data:  x
s = 8, p-value = 0.1094
alternative hypothesis: true median is not equal to 200
95 percent confidence interval:
 199.4867 203.6027
sample estimates:
median of x
     201.75

                  Conf.Level   L.E.pt   U.E.pt
Lower Achieved CI     0.8906 200.5000 203.4000
Interpolated CI       0.9500 199.4867 203.6027
Upper Achieved CI     0.9785 199.0000 203.7000
```

SAS Code

```
data thermo;
input temp @@;
scaletemp = temp - 200;
datalines;
202.2 203.4 200.5 202.5 206.3 198.0 203.7 200.8 201.3 199.0
;
run;

proc univariate data=thermo;
var scaletemp;
run;
```

SAS Output

The UNIVARIATE Procedure
Variable: scaletemp

Moments

N	10	Sum Weights	10
Mean	1.77	Sum Observations	17.7
Std Deviation	2.41018671	Variance	5.809
Skewness	0.2683042	Kurtosis	0.232023
Uncorrected SS	83.61	Corrected SS	52.281
Coeff Variation	136.168741	Std Error Mean	0.76216796

Basic Statistical Measures

Location		Variability	
Mean	1.770000	Std Deviation	2.41019
Median	1.750000	Variance	5.80900
Mode	.	Range	8.30000
		Interquartile Range	2.90000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	2.322323	Pr > |t|	0.0453
Sign	M	3	Pr >= |M|	0.1094
Signed Rank	S	19.5	Pr >= |S|	0.0488

Sign Test: Large Samples ($n \geq 20$)

- If n is large, we can use the CLT for the sign test.
- Recall for $S_+ \sim Bin(n, p)$

$$E(S_+) = np \ , \ Var(S_+) = np(1-p)$$

- Thus

$$Z = \frac{S_+ - np}{\sqrt{np(1-p)}}$$

  will be approx $N(0, 1)$
- The approx gets better as $n \to \infty$

Sign Test: Large Samples

- For sign test, $H_0 : p = 0.5$
- Therefore we compute

$$Z = \frac{S_+ - n/2}{\sqrt{n/4}}$$

- Critical region comes from $\Phi(z)$

Binomial Continuity Correction

- Suppose $S_+ \sim Bin(n, p)$ and $X \sim N(np, np(1-p))$
- By CLT

$$\Pr[S_+ \leq x] \approx \Pr[X \leq x]$$

- Continuity correction

$$\Pr[S_+ \leq x] \approx \Pr[X \leq x + 1/2]$$

- Eg $n = 12$, $p = 0.5$, $x = 4$

$$\Pr[S_+ \leq 4] \approx \Pr[X \leq 4 + 1/2] = 0.193$$

Binomial Continuity Correction

- Likewise

$$\Pr[S_+ \geq x] \approx \Pr[X \geq x - 1/2]$$

Sign Test for Matched Pairs

- Signed test can be applied to matched pair data for hypothesis concerning the population median of the differences.
- In this case S_+ counts the number of positive differences.

Sign Test Summary

Hypothesis	p-value
$H_0 : \tilde{\mu} \leq \tilde{\mu}_0$ $H_1 : \tilde{\mu} > \tilde{\mu}_0$	$P(S_+ \geq s_+) = \sum_{i=s_+}^{n} \binom{n}{i} 0.5^n$
$H_0 : \tilde{\mu} \geq \tilde{\mu}_0$ $H_1 : \tilde{\mu} < \tilde{\mu}_0$	$P(S_- \geq s_-) = \sum_{i=s_-}^{n} \binom{n}{i} 0.5^n$
$H_0 : \tilde{\mu} = \tilde{\mu}_0$ $H_1 : \tilde{\mu} \neq \tilde{\mu}_0$	$2 \sum_{i=s_{\max}}^{n} \binom{n}{i} 0.5^n$

CI based on sign test

- A $(1-\alpha)\%$ CI for $\tilde{\mu}$ is given by

$$x_{(b+1)} \leq \tilde{\mu} \leq x_{(n-b)}$$

where b is the lower $\alpha/2$ critical point of $Bin(n, 0.5)$.

CI based on sign test

- For the thermostat example with $n = 10$ and $p = 0.5$, $\sum_{i=0}^{1} \frac{1}{2^{10}} \binom{10}{i} = 0.0107$ whereas $\sum_{i=0}^{2} \frac{1}{2^{10}} \binom{10}{i} = 0.055$.
- Because of the discreteness of Binomial distribution, we cannot find exact 95% CI for $\tilde{\mu}$.
- A 97.8% CI for $\tilde{\mu}$ is

$$[x_{(2)}, x_{(9)}]$$

- A 89% CI for $\tilde{\mu}$ is

$$[x_{(3)}, x_{(8)}]$$

```
> pbinom(0:3,10,0.5)
 [1] 0.00098 0.0107 0.0547 0.172
```

Wilcoxon Signed Rank Test

- Suppose $Y_1, Y_2, \ldots, Y_n$ iid according to a symmetric distribution F with median $\tilde{\mu}$
- Hypotheses

$$H_0 : \tilde{\mu} = \tilde{\mu}_0$$

vs

$$H_a : \tilde{\mu} > \tilde{\mu}_0$$

Wilcoxon Signed Rank Test

- Delete Y_i's equal $\tilde{\mu}_0$, adjust n
- Compute $Y_i' = Y_i - \tilde{\mu}_0$
- Rank $|Y_i'|$'s from smallest to largest
- The statistic W_+ is the sum of ranks from observation with Y_i' positive
- W_- defined similarly

Wilcoxon Signed Rank Test

- Reject H_0 if W_+ is large or equivalently W_- is small.
- Critical region

$$C_\alpha = \{w_+ : w_+ \geq w_{n,\alpha}\}$$

where

$$P(W_+ \geq w_{n,\alpha} | H_0) \leq \alpha$$

- P-value $= P(W_+ \geq w_+)$.

Wilcoxon Signed Rank Test

- One tailed: $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs $H_a : \tilde{\mu} > \tilde{\mu}_0$
  P-value $= P(W_+ \geq w_+)$
- One tailed: $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs $H_a : \tilde{\mu} < \tilde{\mu}_0$
  P-value $= P(W_- \geq w_-)$
- Two tailed: $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs $H_a : \tilde{\mu} \neq \tilde{\mu}_0$
  P-value $= 2P(W_+ \geq w_{\max})$ where $w_{\max} = \max(w_+, w_-)$

Wilcoxon Signed Rank Test: Example

Example: Thermostat (Example 14.4 of your textbook)

	x	d	Rank	signedRank
[1,]	202.2	2.2	6	6
[2,]	203.4	3.4	8	8
[3,]	200.5	0.5	1	1
[4,]	202.5	2.5	7	7
[5,]	206.3	6.3	10	10
[6,]	198.0	-2.0	5	-5
[7,]	203.7	3.7	9	9
[8,]	200.8	0.8	2	2
[9,]	201.3	1.3	4	4
[10,]	199.0	-1.0	3	-3

Wilcoxon Signed Rank Test: Example

- $W_+ = 6 + 8 + 1 + 7 + 10 + 9 + 2 + 4 = 47$; $W_- = 5 + 3 = 8$
- P-value $= 2P(W_+ \geq 47) = 2 \times 0.024 = 0.048$ from Table A.10
- Therefore, reject H_0 median is 200

R Code and Output

```
> x <- c(202.2, 203.4, 200.5, 202.5, 206.3, 198.0,
203.7, 200.8, 201.3, 199.0)
> d <- x-200
> wilcox.test(d)

Wilcoxon signed rank test

data:  d
V = 47, p-value = 0.04883
alternative hypothesis: true location is not equal to 0

> wilcox.test(x,mu=200)

Wilcoxon signed rank test

data:  x
V = 47, p-value = 0.04883
alternative hypothesis: true location is not equal to 200
```

Distribution of W_+ under H_0

- Large sample distribution
- Can show

$$E(W_+) = \frac{n(n+1)}{4} \text{ and } V(W_+) = \frac{n(n+1)(2n+1)}{24}$$

- If $n \geq 20$,

$$Z = \frac{W_+ - n(n+1)/4 - 0.5}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$$

with continuity correction

Wilcoxon Signed Rank Test: Ties

- If there are 2 or more observations with the same value of Y', the observations are said to be *tied*
- For tied observations we assign the average rank or *midrank*
- Example: $\mathbf{Y} = \{23, 25, 45, 13, 23, 46\}$
  MidRanks: $\{2.5, 4, 5, 1, 2.5, 6\}$

Wilcoxon Signed Rank Test: Ties

- Can show
$$E(W_+) = \frac{n(n+1)}{4}$$
- To accommodate ties, var is adjusted
$$V(W_+) = \frac{n(n+1)(2n+1) - \frac{1}{2}\sum_{i=1}^{q} t_i(t_i-1)(t_i+1)}{24}$$
where q equals the number of sets of ties and t_i is the number of observations in the ith set
- For example on previous slide, $q = 1$ and $t_1 = 2$ such that
$$V(W_+) = \frac{6(6+1)(2\cdot 6+1) - \frac{1}{2}\cdot 2\cdot 1\cdot 3}{24}$$

Wilcoxon Signed Rank Test: Ties

- If n is large, use the variance adjusted for ties in the normal approximation
- If n is small and there are ties, need to compute the null distribution from permutation principles, i.e., tables of critical values are not guaranteed to be correct in the presence of ties

Example 14.5 (Tamhane and Dunlop): Cardiac Output

			D	Rank	SignedRank
[1,]	6.3	5.2	1.1	19.5	19.5
[2,]	6.3	6.6	-0.3	8.0	-8.0
[3,]	3.5	2.3	1.2	21.0	21.0
[4,]	5.1	4.4	0.7	16.0	16.0
[5,]	5.5	4.1	1.4	23.0	23.0
[6,]	7.7	6.4	1.3	22.0	22.0
[7,]	6.3	5.7	0.6	14.0	14.0
[8,]	2.8	2.3	0.5	12.0	12.0
[9,]	3.4	3.2	0.2	5.5	5.5
[10,]	5.7	5.2	0.5	12.0	12.0
[11,]	5.6	4.9	0.7	16.0	16.0
[12,]	6.2	6.1	0.1	2.5	2.5
[13,]	6.6	6.3	0.3	8.0	8.0
[14,]	7.7	7.4	0.3	8.0	8.0
[15,]	5.6	4.9	0.7	16.0	16.0
[16,]	6.3	5.4	0.9	18.0	18.0
[17,]	5.6	5.1	0.5	12.0	12.0
[18,]	4.8	4.4	0.4	10.0	10.0
[19,]	4.2	4.1	0.1	2.5	2.5
[20,]	3.3	2.2	1.1	19.5	19.5
[21,]	3.8	4.0	-0.2	5.5	-5.5
[22,]	5.7	5.8	-0.1	2.5	-2.5
[23,]	4.1	4.0	0.1	2.5	2.5

$n=23$.

4 ties 1. midrange $= \frac{1+2+3+4}{4} = 2.5$

2 ties 2. (1.1) 19.5

3 ties 3. (0.3) 8.

3 ties 4. (0.5) 12

3 ties 5 (0.1) 16

$q=5, t_1=4, t_2=2, t_3=t_4=t_5=3$

$Var(W_+) = \frac{23(23+1)(2\cdot 23+1)}{24} - \frac{1}{2}\left[\frac{143}{}+\frac{12}{}+\frac{13}{}+\frac{}{}+\frac{}{}\right]$

Wilcoxon Signed Rank Test: Example w/ Ties

- $W_+ = 260.5$; $E(W_+) = 23(24)/4 = 138$
- $q = 6$; $t_1 = t_2 = t_6 = 2$; $t_3 = t_4 = t_5 = 3$ such that

$$V(W_+) = \frac{23(24)(47) - \frac{1}{2}\{3(3)(2)(4) + 3(2)(1)(3)\}}{24} = 1079.125$$

- Thus

$$Z = \frac{260.5 - 138 - 0.5}{\sqrt{1079.125}} = 3.714$$

- Yielding $p = 2[1 - \Phi(3.714)] = 0.0002041$

R Code and Output

```
> MethodA <- c(6.3,6.3,3.5,5.1,5.5,7.7, 6.3, 2.8,3.4,
5.7,5.6,6.2,6.6,7.7,7.4,5.6,6.3,8.4,5.6,4.8,4.3,4.2,3.3,3.8,5.7,4.1)
> MethodB <- c(5.2,6.6,2.3,4.4,4.1,6.4,5.7,2.3,3.2,5.2,4.9,6.1,6.3,
7.4,7.4,4.9,5.4,8.4,5.1,4.4,4.3,4.1,2.2,4.0,5.8,4.0)
> wilcox.test(MethodA,MethodB,exact=FALSE,correct=TRUE,paired=TRUE)

Wilcoxon signed rank test with continuity correction

data:  MethodA and MethodB
V = 260.5, p-value = 0.0002041
alternative hypothesis: true location shift is not equal to 0
```