Non-parametric tests

B.H. Robbins Scholars Series

June 24, 2010

Outline

One Sample Test: Wilcoxon Signed-Rank

Two Sample Test: Wilcoxon–Mann–Whitney

Confidence Intervals

Summary

Introduction

- T-tests: tests for the means of continuous data
  - One sample $H_0: \mu = \mu_0$ versus $H_A: \mu \neq \mu_0$
  - Two sample $H_0: \mu_1 - \mu_2 = 0$ versus $H_A: \mu_1 - \mu_2 \neq 0$
- Underlying these tests is the assumption that the data arise from a normal distribution
- T-tests do not actually require normally distributed data to perform reasonably well in most circumstances
- Parametric methods: assume the data arise from a distribution described by a few parameters (Normal distribution with mean μ and variance σ^2).
- Nonparametric methods: do not make parametric assumptions (most often based on ranks as opposed to raw values)
- We discuss non-parametric alternatives to the one and two sample t-tests.

Examples of when the parametric t-test goes wrong

- Extreme outliers
  - Example: t-test comparing two sets of measurements
    - Sample 1: 1 2 3 4 5 6 7 8 9 10
    - Sample 2: 7 8 9 10 11 12 13 14 15 16 17 18 19 20
  - Sample averages: 5.5 and 13.5, T-test p-value $p = 0.000019$
  - Example: t-test comparing two sets of measurements
    - Sample 1: 1 2 3 4 5 6 7 8 9 10
    - Sample 2: 7 8 9 10 11 12 13 14 15 16 17 18 19 20 **200**
  - Sample averages: 5.5 and 25.9, T-test p-value $p = 0.12$

Examples of when the parametric t-test goes wrong

- T-statistic

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

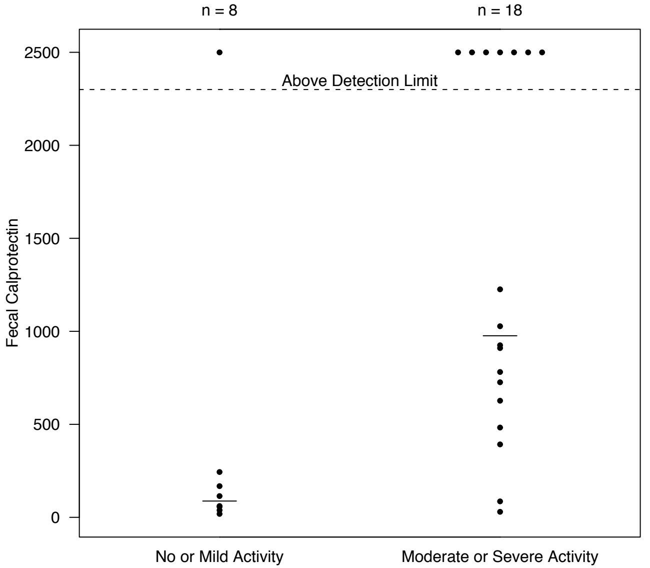
- For two sample tests

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- In the first dataset
  - $s_1^2 = 9.2,\ s_2^2 = 17.5$
- In the second dataset
  - $s_1^2 = 9.2,\ s_2^2 = 2335$

Examples of when the parametric t-test goes wrong

- Upper detection limits
  - Example: Fecal calprotectin was being evaluated as a possible biomarker of Crohn's disease severity
  - Median can be calculated (mean cannot)


n = 8
n = 18
2500
2000
1500
1000
500
0
Fecal Calprotectin
Above Detection Limit
No or Mild Activity
Moderate or Severe Activity

When to use non-parametric methods

- With correct assumptions (e.g., normal distribution), parametric methods will be more efficient / powerful than non-parametric methods but often not as much as you might think[1]
- If the normality assumption grossly violated, nonparametric tests can be much more efficient and powerful than the corresponding parametric test
- Non-parametric methods provide a well-foundationed way to deal with circumstance in which parametric methods perform poorly.

The large-sample efficiency of the Wilcoxon test compared to the t test is $\frac{3}{\pi} = 0.9549$.

Non-parametric methods

- Many non-parametric methods convert raw values to ranks and then analyze ranks
- In case of ties, midranks are used, e.g., if the raw data were 105 120 120 121 the ranks would be 1 2.5 2.5 4

Parametric Test	Nonparametric Counterpart
1-sample t	Wilcoxon signed-rank
2-sample t	Wilcoxon 2-sample rank-sum
k-sample ANOVA	Kruskal-Wallis
Pearson r	Spearman ρ

One sample tests

- Non-parametric analogue to the one sample t-test.
- Almost always used on paired data where the column of values represents differences (e.g., $D = Y_{post} - Y_{pre}$).
- *Sign test:* the simplest test for the median difference being zero in the population
  - Examine all values of D after discarding those in which D=0
  - Count the number of positive Ds
  - Tests H_0 :Prob$[D > 0] = \frac{1}{2}$ versus H_A :Prob$[D > 0] \neq \frac{1}{2}$
    - Under H_0 it is equally likely in the population to have a value below zero as it is to have a value above zero
  - Note that it ignores magnitudes completely → it is inefficient (low power)

One sample tests: Wilcoxon signed rank

- In the pre-post analysis
  - D = pre - post
  - Retain the sign of D (+/-)
  - Rank = rank of $|D|$ (absolute value of D)
  - Signed rank, SR = Sign * Rank
  - Base analyses on SR
- Observations with zero differences are ignored
- Example: A pre-post study

Post	Pre	D	Sign	Rank of $|D|$	Signed Rank
3.5	4	0.5	+	1.5	1.5
4.5	4	-0.5	-	1.5	-1.5
4	5	1.0	+	4.0	4.0
3.9	4.6	0.7	+	3.0	3.0

One sample tests

- A good approximation to an exact P-value (not discussed) may be obtained by computing

$$z = \frac{\sum SR_i}{\sqrt{\sum SR_i^2}},$$

where the signed rank for observation i is SR_i.

- We can then compare $|z|$ to the normal distribution.
- Here, $z = \frac{7}{\sqrt{29.5}} = 1.29$ and by `surfstat` the 2-tailed P-value is 0.197
- If all differences are positive or all are negative, the exact 2-tailed P-value is $\frac{1}{2^{n-1}}$
  - This implies that n must exceed 5 for any possibility of significance at the $\alpha = 0.05$ level for a 2-tailed test

One sample tests

- Sleep Dataset
  - Compare the effects of two soporific drugs.
  - Each subject receives Drug 1 and Drug 2
  - Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?
  - Dependent variable: Difference in hours of sleep comparing Drug 2 to Drug 1
  - H_0 : For any given subject, the difference in hours of sleep is equally likely to be positive or negative

Subject	Drug 1	Drug 2	Diff (2-1)	Sign	Rank
1	1.9	0.7	−1.2	-	3
2	−1.6	0.8	2.4	+	8
3	−0.2	1.1	1.3	+	4.5
4	−1.2	0.1	1.3	+	4.5
5	−0.1	−0.1	0.0	NA	NA
6	3.4	4.4	1.0	+	2
7	3.7	5.5	1.8	+	7
8	0.8	1.6	0.8	+	1
9	0.0	4.6	4.6	+	9
10	2.0	3.4	1.4	+	6

Table: Hours of extra sleep on drugs 1 and 2, differences, signs and ranks of sleep study data

One sample / paired test example

- Approximate p-value calculation

$$\sum_{i=1}^{9} SR_i = 39, \quad \sqrt{\sum_{i=1}^{9} SR_i^2} = 16.86$$

  Z = 2.31, and the two sided test yields a p-value equal to 2*(1-.989556)= 0.0209
- Wilcoxon signed rank test statistical program output

```
            Wilcoxon signed rank test
data:  sleep.data
V = 42, p-value = 0.02077
alternative hypothesis: true location is not equal to 0
```

- Thus, we reject H_0 and conclude Drug 2 increases sleep by more hours than Drug 1 ($p = 0.02$)

One sample / paired test example

- We could also perform sign test on sleep data
  - If drugs are equally effective, we should have same number of positives and negatives (e.g., Prob(D>0)=.5).
  - Analogous to coin flip example from last time.
  - In the observed data: 1 negative and 8 positives (we throw out 1 'no change')
  - One sided p-value: probability of observing 0 or 1 negatives
  - Two sided p-value: probability of observing 0, 1, 8, or 9 negatives
  - $p = 0.04$, $\rightarrow$ reject H_0 at $\alpha = 0.05$

Wilcoxon signed rank test

- Assumes the distribution of differences is symmetric
- When the distribution is symmetric, the signed rank test tests whether the median difference is zero
- In general it tests that, for two randomly chosen observations i and j with values (differences) x_i and x_j, that the probability that $x_i + x_j > 0$ is $\frac{1}{2}$
- The estimator that corresponds exactly to the test in all situations is the pseudomedian, the median of all possible pairwise averages of x_i and x_j, so one could say that the signed rank test tests H_0: pseudomedian=0

- To test $H_0 : \eta = \eta_0$, where η is the population median (not a difference) and η_0 is some constant, we create the n values $x_i - \eta_0$ and feed those to the signed rank test, assuming the distribution is symmetric
- When all nonzero values are of the same sign, the test reduces to the *sign test* and the 2-tailed P-value is $(\frac{1}{2})^{n-1}$ where n is the number of nonzero values

Two sample WMW test

- The Wilcoxon–Mann–Whitney (WMW) 2-sample rank sum test is for testing for equality of central tendency of two distributions (for unpaired data)
- Ranking is done by combining the two samples and ignoring which sample each observation came from
- Example:

Females	120	118	121	119
Males	124	120	133	
Ranks for Females	3.5	1	5	2
Ranks for Males	6	3.5	7	

Two sample WMW test

- Doing a 2-sample *t*-test using these ranks as if they were raw data and computing the *P*-value against 4+3-2=5 d.f. will work quite well
- Loosely speaking the WMW test tests whether the population medians of the two groups are the same
- More accurately and more generally, it tests whether observations in one population tend to be larger than observations in the other
- Letting x_1 and x_2 respectively be randomly chosen observations from populations one and two, WMW tests $H_0 : C = \frac{1}{2}$, where $C = \text{Prob}[x_1 > x_2]$

Two sample WMW test

- Wilcoxon rank sum test statistic

$$W = R - \frac{n_1(n_1+1)}{2}$$

where R is the sum of the ranks in group 1

- Under H_0, $\mu_w = \frac{n_1 n_2}{2}$ and $\sigma_w = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$, and

$$z = \frac{W - \mu_w}{\sigma_w}$$

follow a N(0,1) distribution.

Two sample WMW test

- The C index (*concordance probability*) may be estimated by computing

$$C = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2},$$

where $\bar{R}$ is the mean of the ranks in group 1

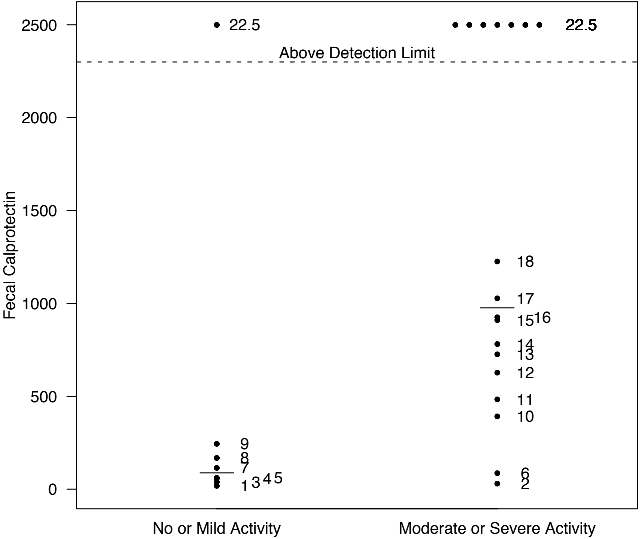
- For the above data $\bar{R} = 2.875$ and $C = \frac{2.875-2.5}{3} = 0.125$
- We estimate: probability that a randomly chosen female has a value greater than a randomly chosen male is 0.125.

Two sample WMW test: Example

- Fecal calprotectin being evaluated as a possible biomarker of disease severity
- Calprotectin measured in 26 subjects, 8 observed to have no/mild activity by endoscopy
- Calprotectin has upper detection limit at 2500 units
    - A type of missing data, but need to keep in analysis

Two sample WMW test: Example

- Study question: Are calprotectin levels different in subjects with no or mild activity compared to subjects with moderate or severe activity?
- Statement of the null hypothesis
  - H_0 : Populations with no/mild activity have the same distribution of calprotectin as populations with moderate/severe activity
  - $H_0 : C = \frac{1}{2}$


Fecal Calprotectin
2500
2000
1500
1000
500
0
22.5
22.5
Above Detection Limit
18
17
15
16
14
13
12
11
10
9
8
7
1
3
4
5
6
2
No or Mild Activity
Moderate or Severe Activity

Two sample WMW test: Example

- Stat program output

```
          Wilcoxon rank sum test
data:  calpro by endo2
W = 23.5, p-value = 0.006257
alternative hypothesis: true location shift is not equa
```

- $W = 59.5 - \frac{8*9}{2} = 23.5$
- A common (but loose) interpretation: People with moderate/severe activity have higher *median* fecal calprotectin levels than people with no/mild activity ($p = 0.006$).

Confidence Intervals for medians

- Confidence intervals for the (one sample) median
  - Ranks of the observations are used to give approximate confidence intervals for the median (See Altman book)
  - e.g., if $n = 12$, the 3^{rd} and 10^{th} largest values give a 96.1% confidence interval
  - For larger sample sizes, the lower ranked value (r) and upper ranked value (s) to select for an approximate 95% confidence interval for the population median is

$$r = \frac{n}{2} - 1.96 * \frac{\sqrt{n}}{2} \quad \text{and} \quad s = 1 + \frac{n}{2} + 1.96 * \frac{\sqrt{n}}{2}$$

  - e.g., if $n = 100$ then $r = 40.2$ and $s = 60.8$, so we would pick the 40^{th} and 61^{st} largest values from the sample to specify a 95% confidence interval for the population median

Confidence Intervals

- Confidence intervals for the difference in two medians (two samples)
  - Considers all possible differences between sample 1 and sample 2

	Female			
Male	120	118	121	119
124	4	6	3	5
120	0	2	-1	1
133	13	15	12	14

- An estimate of the median difference (males - females) is the median of these 12 differences, with the 3^{rd} and 10^{th} largest values giving an (approximate) 95% CI
- Median estimate = 4.5, 95% CI = [1, 13]
- Specific formulas found in Altman, pages 40-41

Confidence Intervals

- Bootstrap
  - General method, not just for medians
  - Non-parametric, does not assume symmetry
  - Iterative method that repeatedly samples from the original data
  - Algorithm for creating a 95% CI for the difference in two medians
    1. Sample *with replacement* from sample 1 and sample 2
    2. Calculate the difference in medians, save result
    3. Repeat Steps 1 and 2 1000 times
  - A (naive) 95% CI is given by the 25^{th} and 97.5^{th} largest values of your 1000 median differences
  - For the male/female data, median estimate = 4.5, 95% CI = [-0.5, 14.5], which agrees with the conclusion from a WMW rank sum test ($p = 0.11$).

Summary: non-parametric tests

- Wilcoxon signed rank test: alternative to the one sample t-test
- Wilcoxon Mann Whitney or rank sum test: alternative to the two sample t-test
- Attractive when parametric assumptions are believed to be violated
- Drawback: if based on ranks, tests do not provide insight into effect size
- Non-parametric tests are attractive if all we care about is getting a P-value