MET3220C
Computational Statistics

Hypothesis Testing:
Classical Nonparametric tests

(Chapter 5.3 of Wilks' book)

Key Points:
1) Tests for location

Nonparametric Tests

- Some statistical tests do not require the assumption statistical distribution or a null distribution. That is, they don't require knowledge of:
  - The sampling distribution of the data, which is used to calculate probabilities. This is the null distribution.
    - Example: the number of hurricane landfalls is a Poisson distribution.
    - Example: differences in a mean follow a Gaussian distribution.
- Nonparametric test are appropriate if either or both of the following conditions apply:
  - We know or suspect that the parametric assumption(s) required for a particular test are not met.
    - Example: grossly non-Gaussian data for a t test.
  - A test statistic is a complicated function of the data, and its sampling distribution is unknown or cannot be derived.

Differences Between Parametric and Nonparametric Tests

- The big difference between these two types of tests is the approach used to determine the null distribution.
  - That is, step 4 on the list of 5 steps.
- The other steps are similar between these two types of tests.
  - 1) Identify a test statistic
    - Chose a statistic that is appropriate to the data and the question.
  - 2) Define a null hypothesis.
    - The null hypothesis (H_0) defines a logical structure which will be used to examine the test statistic.
    - The null hypothesis is often designed as the compliment to what we would like to test for.
  - 3) Define the alternative hypothesis (H_A).
  - 4) Determine the null distribution.
  - 4.5) Chose a confidence limit.
  - 5) Compare the test statistic to the null distribution.

Two Types of Nonparametric Tests

- Classical nonparametric testing is based on a mathematical analysis related to the null hypothesis.
  - Note: saying that the test is related to the null hypothesis also means that the test is related to the alternative hypothesis.
  - Classical techniques were usually developed prior to efficient computing power.
  - References:
    - Conover, 1999: Practical Nonparametric Statistics, Wiley, 584pp.
    - Daniel, 1990: Applied Nonparametric Statistics. Kent, 635 pp.
    - Sprent and Smeeton, 2001: Applied Nonparametric Statistical Methods. Chapman and Hall, 461pp.
- Resampling techniques
  - The null distribution is determined empirically, based on the available data.
  - Applicable to just about any test statistic.

Classical Tests for Location
Wilcoxon-Mann-Whitney Rank-Sum Test

- Location is a nonparametric analog to the mean.
- Wilcoxon-Mann-Whitney rank-sum test
  - Independently discovered in the 1940s by Wilcoxon as well as Mann and Whitney.
  - Applies to two independent (and non-paired) samples.
  - The null hypothesis is that the two data samples have been drawn from the same distribution.
- One sided and two sided hypotheses are possible.
- The effect of serial correlation is similar to the effect on the t test.

Wilcoxon-Mann-Whitney Rank-Sum Test
The Concept

- If the null hypothesis is true:
  - The two data samples are drawn from the same distribution.
  - The labeling of each datum value as belonging to one group or the other is arbitrary.
- The null hypothesis is equivalent to saying that rather than two samples, made up of n_1 and n_2 data points, there is one sample made up of $n = n_1 + n_2$ data points.
  - The concept that the labels are arbitrary, because they have been drawn from the sample distribution, is known as the principle of exchangeability.
- The rank sum test statistic is not a function of the data values.
  - It is a function of the ranks within the n pooled samples.
  - This approach makes the data distribution irrelevant.

Wilcoxon-Mann-Whitney Rank-Sum Test
The Details

- R_1 is the sum of ranks in sample 1, and R_2 is the sum of ranks in sample 2.
  - $R_1 + R_2 = 1 + 2 + 3 + \ldots + n = n (n - 1) / 2$
  - If H_0 is true, and if $n_1 = n_2$, then R_1 should be similar to R_2.
  - If $n_1 \neq n_2$, then R_1 / n_1 should be similar to R_2 / n_2.
- If the null hypothesis is true, then there are many equally likely ways the data could be partitioned into groups of size n_1 and n_2.
  - Specifically, there are $(n!) / [(n_1!) (n_2!)]$ equally likely arrangements.
  - For example, with $n_1 = n_2 = 10$, there are 184,756 arrangements.
- It is not necessary to compute R_1 and R_2 for this vast number of combinations.
  - Instead, use the Mann-Whitney U-statistic:
    - $U_1 = R_1 - 0.5\, n_1 (n_1+1)$, or
    - $U_2 = R_2 - 0.5\, n_2 (n_2+1)$.

Wilcoxon-Mann-Whitney Rank-Sum Test
More Details

- Only one of the Mann-Whitney U-statistics is needed because any one of them can be determined from the other.
  - $(U_1 + U_2) = n_1 n_2$
- If there are 10 or more data in each sample, the U-statistic is approximately Gaussian.
  - The mean $\mu_U = n_1 n_2 / 2$, and
  - Standard deviation $\sigma_U = \left[\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right]^{1/2}$
- For smaller samples, tables of critical values (e.g., Conover 1999) can be used.
- Note: if there are multiple occurrences of an outcome, these can all be treated as the same rank, with that rank equal to the average of the ranks occupied by the like observations.

Example: Does Cloud Seeding Influence Lightning Strikes?

TABLE 5.5 Counts of cloud-to-ground lightning for experimentally seeded and nonseeded storms. From Baughman *et al.* (1976).

Seeded		Unseeded	
Date	Lightning strikes	Date	Lightning strikes
7/20/65	49	7/2/65	61
7/21/65	4	7/4/65	33
7/29/65	18	7/4/65	62
8/27/65	26	7/8/65	45
7/6/66	29	8/19/65	0
7/14/66	9	8/19/65	30
7/14/66	16	7/12/66	82
7/14/66	12	8/4/66	10
7/15/66	2	9/7/66	20
7/15/66	22	9/12/66	358
8/29/66	10	7/3/67	63
8/29/66	34		

Table from Wilks' Statistical Methods for Atmospheric Sciences

- It was suspected that seeding clouds would reduce lightning.
- An experiment was designed to seed, or not seed, clouds with similar characteristics, and record lightning characteristics.
- We will examine ground strikes.
- There were 12 seed storms (n_1), with a mean of 19.25 strikes.
- There were 11 unseeded storms (n_2), with an averages of 69.45 strikes.

Example: Does Cloud Seeding Influence Lightning Strikes?

TABLE 5.6 Illustration of the procedure of the rank-sum test using the cloud-to-ground lightning data in Table 5.5. In the left portion of this table, the $n_1 + n_2 = 23$ counts of lightning strikes are pooled and ranked. In the right portion of the table, the observations are segregated according to their labels of seeded (S) or not seeded (N) and the sums of the ranks for the two categories (R_1 and R_2) are computed.

Pooled Data			Segregated Data			
Strikes	Seeded?	Rank				
0	N	1			N	1
2	S	2	S	2		
4	S	3	S	3		
9	S	4	S	4		
10	N	5.5			N	5.5
10	S	5.5	S	5.5		
12	S	7	S	7		
16	S	8	S	8		
18	S	9	S	9		
20	N	10			N	10
22	S	11	S	11		
26	S	12	S	12		
29	S	13	S	13		
30	N	14			N	14
33	N	15			N	15
34	S	16	S	16		
45	N	17			N	17
49	S	18	S	18		
61	N	19			N	19
62	N	20			N	20
63	N	21			N	21
82	N	22			N	22
358	N	23			N	23
		Sums of Ranks:	$R_1 = 108.5$		$R_2 = 167.5$	

- There is one huge outlier in the unseeded data set. This occurrence suggests that it will be difficult to fit the data to parametric distributions and have an acceptably small uncertainty in the fitting parameters.
  - Therefore, a non-parametric test is appropriate.
  - Note: the standard deviation of the unseeded sample is 98.93 strikes.
- $R_1 = 108.5$ and $R_2 = 167.5$
- $U_1 = 108.5 - 6\ (12 + 1) = 30.5$
- $\mu_U = (12)\ (11) / 2 = 66$
- $\sigma_U = [(12)\ (11)\ (12+11+1)/12]^{1/2} = 16.2$
- $z = (30.5 - 66) / 16.2 = -2.19$
- Results in a one tailed probability of 0.014.

Wilcoxon Signed Rank Test
(for paired data from two samples)

- This test takes advantage of the positive correlation between the sets of paired data.
  - Example: comparison of time series of like observations from different locations (but for the same times).
- Notation:
  - n pairs of data: (x_i, y_i), for $i = 1, n$.
  - $D_i = x_i - y_i$
- Null Hypothesis: The data (x and y) are drawn from the same population.
  - The number of positive values of D should be similar to the number of negative values of D.
- Transform $|D_i|$ to ranks:
  - $T_i = \text{rank}(|D_i|)$
  - Values of $|D_i| = 0$ will be ignored in subsequent tests.

Wilcoxon Signed Rank Test
The Details

- Now separately sum the ranks for positive and negative values of D:

$$T^+ = \sum_{D_i > 0} T_i$$

$$T^- = \sum_{D_i < 0} T_i$$

- Let n' be the number of non-zero values of D_i.
  - We then know that $T^+ + T^- = n'(n'+1)/2$
- Recall that the null hypothesis is equivalent to saying that assigning assigning one of the values in a pair to the x sample is arbitrary.
  - If so, there are $2^{n'}$ equally likely arrangements of 2n' data values.
  - The null distribution is based on these $2^{n'}$ values of T.
  - For $n' > 20$ the distribution is well approximated as Gaussian.
  - For smaller values of n', tables of critical values exist (e.g., Conover 1999).

Wilcoxon Signed Rank Test
The Details

- If there are sufficient pairs for the test distribution to be Gaussian, then

$$\mu_T = \frac{n'(n'+1)}{4}$$

Note $\mu_T = (T^+ + T^-)/2$

$$\sigma_T = \left[\frac{n'(n'+1)(2n'+1)}{24}\right]^{0.5}$$

- z-scores can be estimated by taking the difference from the mean, and dividing this difference by the standard deviation.

$$z = \frac{(T^+ - \mu_T)}{\sigma_T} = -\frac{(T^- - \mu_T)}{\sigma_T}$$

Wilcoxon Signed Rank Test
Example: Thunderstorm Frequency

TABLE 5.7 Illustration of the procedure of the Wilcoxon signed-rank test using data for counts of thunderstorms reported in the northeastern United States (x) and the Great Lakes states (y) for the period 1885–1905, from Brooks and Carruthers (1953). Analogously to the procedure of the rank-sum test (see Table 5.6), the absolute values of the annual differences, $|D_i|$, are ranked and then segregated according to whether D_i is positive or negative. The sum of the ranks of the segregated data constitute the test statistic.

Paired Data			Differences		Segregated Ranks	
Year	X	Y	D_i	Rank$\lvert D_i \rvert$	$D_i > 0$	$D_i < 0$
1885	53	70	−17	20		20
1886	54	66	−12	17.5		17.5
1887	48	82	−34	21		21
1888	46	58	−12	17.5		17.5
1889	67	78	−11	16		16
1890	75	78	−3	4.5		4.5
1891	66	76	−10	14.5		14.5
1892	76	70	+6	9	9	
1893	63	73	−10	14.5		14.5
1894	67	59	+8	11.5	11.5	
1895	75	77	−2	2		2
1896	62	65	−3	4.5		4.5
1897	92	86	+6	9	9	
1898	78	81	−3	4.5		4.5
1899	92	96	−4	7		7
1900	74	73	+1	1	1	
1901	91	97	−6	9		9
1902	88	75	+13	19	19	
1903	100	92	+8	11.5	11.5	
1904	99	96	+3	4.5	4.5	
1905	107	98	+9	13	13	
				Sums of Ranks:	$T^+ = 78.5$	$T^- = 152.5$

reported thunderstorms, the test is two-tailed (H_A is simply "not H_0"), so the p valu[e] $\Pr\{z \leq -1.29\} + \Pr\{z > +1.29\} = 2\Pr\{z \leq -1.29\} = 0.197$. The null hypothesis wo[uld] not be rejected in this case. Note that the same result would be obtained if the test stati[stic] $T^- = 152.5$ had been chosen instead. ◊

- Examine annual thunderstorm numbers in the northeastern US (x) and in the Great Lakes states (y).
  - 21 years of observations
  - 1885 to 1905
- Synoptic conditions are similar in these regions, therefore the annual number of storms might be expected to be similar.
- No values of D equal zero: n' = 21.
- T^+ = 78.5, and T^- = 152.5
- μ_T = (21) (22)/4 = 115.5
  σ_T = [(21) (22) (43)/24]$^{1/2}$ = 28.77
- z = (152.5 − 115.5) / 28.77 = 1.29
- $2\Pr\{z \geq 1.29\}$ = 0.197