UCLA STAT 110B

Applied Statistics for Engineering and the Sciences

•**Instructor**: **Ivo Dinov**,
Asst. Prof. In Statistics and Neurology

•**Teaching Assistants:** Brian Ng, UCLA Statistics

University of California, Los Angeles, Spring 2003
http://www.stat.ucla.edu/~dinov/courses_students.html

Why Use Nonparametric Statistics?

- Parametric tests are based upon assumptions that may include the following:
  - The data have the **same variance**, regardless of the treatments or conditions in the experiment.
  - The data are **normally distributed** for each of the treatments or conditions in the experiment.
- What happens when we are not sure that these assumptions have been satisfied?

How Do Nonparametric Tests Compare with the Usual *z*, *t*, and *F* Tests?

- Studies have shown that when the usual assumptions are satisfied, nonparametric tests are about 95% efficient when compared to their parametric equivalents.
- When normality and common variance are not satisfied, the nonparametric procedures can be much more efficient than their parametric equivalents.

The Wilcoxon Rank Sum Test

- Suppose we wish to test the hypothesis that two distributions have the same center.
- We select two independent random samples from each population. Designate each of the observations from population 1 as an "**A**" and each of the observations from population 2 as a "**B**".
- If H_0 is true, and the two samples have been drawn from the same population, when we rank the values in both samples from small to large, **the A's and B's should be randomly mixed in the rankings**.

What happens when H_0 is true?

•Suppose we had 5 measurements from population 1 and 6 measurements from population 2.

•If they were drawn from the same population, the rankings might be like this.

ABABBBABABBA

•In this case if we summed the ranks of the A measurements and the ranks of the B measurements, the sums would be similar.

What happens if H_0 is not true?

- If the observations come from two different populations, perhaps with **population 1 lying to the left of population 2**, the ranking of the observations might take the following ordering.

AAABABABBB

In this case the sum of the ranks of the B observations would be larger than that for the A observations.

How to Implement Wilcoxon's Rank Test

- Rank the combined sample from smallest to largest.
- Let T_1 represent the sum of the ranks of the first sample (**A**'s).
- Then, T_1^* defined below, is the sum of the ranks that the A's would have had if the observations were ranked from *large to small.*

$$T_1^* = n_1(n_1 + n_2 + 1) - T_1$$

The Wilcoxon Rank Sum Test

H_0: the two population distributions are the same

H_a: the two populations are in some way different

- The **test statistic** is the smaller of T_1 and T_1^*.
- Reject H_0 if the test statistic is less than the **critical value** found in Table 7(a).
- Table 7(a) is indexed by letting population 1 be the one associated with the smaller sample size n_1, and population 2 as the one associated with n_2, the larger sample size.

Example

The wing stroke frequencies of two sp... for a sample of n_1 = ... from species 2. C...tributions of wing strokes differ for these two species? Use α = .05

If several measurements are tied, each gets the average of the ranks they would have gotten, if they were not tied! (See x = 180)

H_0: the two species are the same

H_a: the two species are in some way different

Species 1	Species 2
235 (10)	180 (3.5)
225 (9)	169 (1)
190 (8)	180 (3.5)
188 (7)	185 (6)
	178 (2)
	182 (5)

1. The sample with the smaller sample size is called sample 1.
2. We rank the 10 observations from smallest to largest, shown in parentheses in the table.

The Bee Problem

Can you conclude that the distributions of wing strokes differ for these two species? α = .05.

TABLE 7(b)
2.5% Left-Tailed Critical Values

n_2 \ n_1	2	3	4	5	6	7	8	9
4	—	—	10					
5	—	6	11	17				
6	—	7	12	18	26			
7	—	7	13	20	27	36		
8	3	8	14	21	29	38	49	
9	3	8	14	22	31	40	51	62

...fficient ...nce in ...ke

$\ldots 9 + 10 = 34$

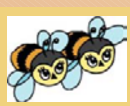
$\ldots + n_2 + 1) - T_1$

$\ldots 6 + 1) - 34 = 10$

	178 (2)
	182 (5)

1. The test statistic is T = 10.
2. The critical value of T from Table 7(b) for a two-tailed test with $\alpha/2$ = .025 is T = 12; H_0 is rejected if T $\leq$ 12.

Minitab Output



Recall $T_1 = 34; T_1^* = 10.$

Mann-Whitney Test and CI: Species1, Species2

```
Species1   N =   4     Median =      207.50
Species2   N =   6     Median =      180.00
Point estimate for ETA1-ETA2 is       30.50
95.7 Percent CI for ETA1-ETA2 is (5.99,56.01)
W = 34.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant
at 0.0142
The test is significant at 0.0139 (adjusted for ties)
```

Minitab calls the procedure the Mann-Whitney U Test, equivalent to the Wilcoxon Rank Sum Test.

The test statistic is W = T_1 = 34 and has *p*-value = .0142. Do not reject H_0 for α = .05.

Large Sample Approximation: Wilcoxon Rank Sum Test

When n_1 and n_2 are large (greater than 10 is large enough), a normal approximation can be used to approximate the critical values in Table 7.

1. Calculate T_1 and T_1^*. Let $T = \min(T_1, T_1^*)$.
2. The statistic $z = \dfrac{T - \mu_T}{\sigma_T}$ has an approximate z distribution with

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \text{ and } \sigma_T^2 = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

$$T_1^* = n_1(n_1 + n_2 + 1) - T_1$$

T_1 = sum of the ranks of sample 1 (**A**'s).

Some Notes

- When should you use the Wilcoxon Rank Sum test instead of the two-sample *t* test for independent samples?
  - ✓ when the responses can only be **ranked** and not quantified (e.g., ordinal qualitative data)
  - ✓ when the F test or the Rule of Thumb shows a **problem with equality of variances**
  - ✓ when a normality plot shows a **violation of the normality assumption**

The Sign Test

- The **sign test** is a fairly simple procedure that can be used to compare two populations when the samples consist of **paired observations.**
- It can be used
  - ✓ when the assumptions required for the **paired-difference test** are not valid or
  - ✓ when the responses can only be ranked as "**one better than the other**", but cannot be quantified.

The Sign Test

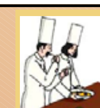
✓ For each pair, measure whether the first response—say, A—exceeds the second response—say, B.

✓ The test statistic is x, the number of times that A exceeds B in the n pairs of observations.

✓ Only pairs without ties are included in the test.

✓ Critical values for the rejection region or exact *p*-values can be found using the cumulative binomial distribution (SOCR resource online).

The Sign Test



H_0: the two populations are identical versus

H_a: one or two-tailed alternative

is equivalent to

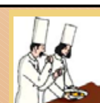
H_0: p = P(A exceeds B) = .5 versus

H_a: p (≠, <, or >) .5

Test statistic: x **= number of plus signs**

Rejection region, *p*-values from Bin(n=size, p).

Example

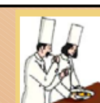


Two gourmet chefs each tasted and rated eight different meals from 1 to 10. Does it appear that one of the chefs tends to give higher ratings than the other? Use α = .01.

Meal	1	2	3	4	5	6	7	8
Chef A	6	4	7	8	2	4	9	7
Chef B	8	5	4	7	3	7	9	8
Sign	-	-	+	+	-	-	0	-

H_0: the rating distributions are the same ($p = .5$)
H_a: the ratings are different ($p \neq .5$)

The Gourmet Chefs



Meal	1	2	3	4	5	6	7	8
Chef A	6	4	7	8	2	4	9	7
Chef B	8	5	4	7	3	7	9	8
Sign	-	-	+	+				

H_0: $p = .5$
H_a: $p \neq .5$ with $n = 7$ (omi
Test Statistic: x = number o

p-value =.454 is too large to reject H_0. There is insufficient evidence to indicate that one chef tends to rate one meal higher than the other.

Use Table 1 with n = 7 and p = .5.
p-value = P(observe x = 2 or something equally as unlikely) = $P(x \leq 2) + P(x \geq 5) = 2(.227) = .454$

k	0	1	2	3	4	5	6	7
$P(x \leq k)$	.008	.062	.227	.500	.773	.938	.992	1.000

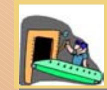
Large Sample Approximation: The Sign Test

When $n \geq 25$, a normal approximation can be used to approximate the critical values of Binomial distribution.

1. Calculate x = number of plus signs.
2. The statistic $z = \dfrac{x - .5n}{.5\sqrt{n}}$ has an approximate z distribution.

Y~Bin(n, p) ➔
E(Y) = np
Var(Y) = np(1-p)

Example



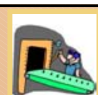
You record the number of accidents per day at a large manufacturing plant for both the day and evening shifts for n = 100 days. You find that the number of accidents per day for the evening shift x_E exceeded th[e ...] in the day shift x_D on 63 of t[he ...] sufficient evidence to indica[te ...] on one shift than on the othe[r ...]

For a two tailed test, we reject H_0 if $|z| > 1.96$ (5% level). H_0 is rejected. There is evidence of a difference between the day and night shifts.

H_0: the distributions (# of a[...]
H_a: the distributions are different ($p \neq .5$)

$$\text{Test statistic: } z = \frac{x - .5n}{.5\sqrt{n}} = \frac{63 - .5(100)}{.5\sqrt{100}} = 2.60$$

Which test should you use?



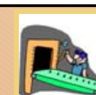
- We compare statistical tests using

  Definition: Power = $1 - \beta$

  = P(reject H_0 when H_a is true)

- The **power** of the test is the probability of rejecting the null hypothesis when it is false and some specified alternative is true.
- The **power** is the probability that the test will do what it was designed to do—that is, detect a departure from the null hypothesis when a departure exists.

Which test should you use?



- If all **parametric assumptions** have been met, the **parametric test will be the most powerful**.
- If not, a nonparametric test may be more powerful.
- If you can reject H_0 with a less powerful nonparametric test, you will not have to worry about parametric assumptions.
- If not, you might try
  - more powerful nonparametric test or
  - increasing the sample size to gain more power

The Wilcoxon Signed-Rank Test – different form Wilcoxon Rank Sum Test

- The **Wilcoxon Signed-Rank Test** is a more powerful nonparametric procedure that can be used to compare two populations when the samples consist of **paired observations.**
- It uses the **ranks** of the differences, $d = x_1 - x_2$ that we used in the paired-difference test.

The Wilcoxon Signed-Rank Test – different form Wilcoxon Rank Sum Test

✓For each pair, calculate the difference $d = x_1 - x_2$. Eliminate zero differences.

✓Rank the absolute values of the differences from 1 to n. Tied observations are assigned average of the ranks they would have gotten if not tied.

- **T^+ = rank sum for positive differences**
- **T^- = rank sum for negative differences**

✓If the two populations are the same, T^+ and T^- should be nearly equal. If either T^+ or T^- is unusually large, this provides evidence against the null hypothesis.

The Wilcoxon Signed-Rank Test

H_0: the two populations are identical versus

H_a: one or two-tailed alternative

Test statistic: $T = \min(T^+ \text{ and } T^-)$

Critical values for a one or two-tailed rejection region can be found using **Wilcoxon Signed-Rank Test Table.**

Example

To compare the densities of cakes using mixes A and B, six pairs of pans (A and B) were baked side-by-side in six different oven locations. Is there evidence of a difference in density for the two cake mixes?

Location	1	2	3	4	5	6
Cake Mix A	.135	.102	.098	.141	.131	.144
Cake Mix B	.129	.120	.112	.152	.135	.163
$d = x_A - x_B$	.006	-.018	-.014	-.011	-.004	-.019

H_0: the density distributions are the same
H_a: the density distributions are different

Cake Densities

Location	1	2	3
Cake Mix A	.135	.102	.098
Cake Mix B	.129	.120	.112
$d = x_A - x_B$	.006	-.018	-.014

Do not reject H_0. There is insufficient evidence to indicate that there is a difference in densities for the two cake mixes.

Rank	2	5	4	3	1	6

Calculate: $T^+ = 2$ and $T^- = 5+4+3+1+6 = 19$.

The **test statistic** is $T = \min(T^+, T^-) = 2$.

Rejection region: Use Table 8. For a two-tailed test with $\alpha = .05$, reject H_0 if $T \le 1$.

TABLE 8
Critical Values of T in the Wilcoxon Signed-Rank Test, $n = 5(1)50$

One-Sided	Two-Sided	$n = 5$	$n = 6$	$n = 7$
$\alpha = .050$	$\alpha = .10$	1	2	4
$\alpha = .025$	$\alpha = .05$		1	2
$\alpha = .010$	$\alpha = .02$			0
$\alpha = .005$	$\alpha = .01$			

Large Sample Approximation: The Signed-Rank Test

When $n \ge 25$, a normal approximation can be used to approximate the critical values in Table 8.

1. Calculate T^+ and T^-. Let $T = \min(T^+, T^-)$.

2. The statistic $z = \dfrac{T - \mu_T}{\sigma_T}$ has an approximate z distribution with

$$\mu_T = \frac{n(n+1)}{4} \text{ and } \sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

The Kruskal-Wallis – *H* Test

- The **Kruskal-Wallis *H* Test** is a nonparametric procedure that can be used to compare more than two populations in a **completely randomized design.**
- **Non-parametric equivalent to ANOVA F-test!**
- All $n = n_1 + n_2 + \ldots + n_k$ measurements are jointly ranked.
- We use the sums of the ranks of the k samples to compare the distributions.

The Kruskal-Wallis – *H* Test

✓Rank the total measurements in all k samples from 1 to n. Tied observations are assigned average of the ranks they would have gotten if not tied.

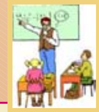
✓Calculate

- **T_i = rank sum for the i^{th} sample $i = 1, 2, \ldots, k$**
- **$n = n_1 + n_2 + \ldots + n_k$**

✓And the test statistic H is (analog to: $F = MSST/MSSE$)

$$H = \frac{12}{n(n+1)} \left(\sum_{i=1}^{k} \frac{T_i^2}{n_i} \right) - 3(n+1)$$

The Kruskal-Wallis *H* Test



H_0: the k distributions are identical versus

H_a: at least one distribution is different

Test statistic: ***Kruskal-Wallis H***

When H_0 is true, the test statistic H has an approximate χ^2 distribution with $df = k$-1.

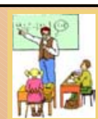
Use a right-tailed rejection region or p-value based on the Chi-square distribution.

Example

Four groups of students were randomly assigned to be taught with four different techniques, and their achievement test scores were recorded. Are the distributions of test scores the same, or do they differ in location?

1	2	3	4
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88

Teaching Methods



	1	2	3	4
	65 (3)	75 (7)	59 (1)	94 (16)
	87 (13)	69 (5)	78 (8)	89 (15)
	73 (6)	83 (12)	67 (4)	80 (10)
	79 (9)	81 (11)	62 (2)	88 (14)
T_i	31	35	15	55

Rank the 16 measurements from 1 to 16, and c... the f... sums

H_0: the distributions of scores are the same
H_a: the distributions differ in location

$$\text{Test statistic: } H = \frac{12}{n(n+1)}\sum\frac{T_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{16(17)}\left(\frac{31^2+35^2+15^2+55^2}{4}\right) - 3(17) = 8.96$$

Teaching Methods

H_0: the distributions of scores are the same
H_a: the distributions differ in location

$$\text{Test statistic: } H = \frac{12}{n(n+1)}\sum\frac{T_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{16(17)}\left(\frac{31^2+35^2+15^2+55^2}{4}\right) - 3(17) = 8.96$$

Rejection region: Use Table 5. For a right-tailed chi-square test with $\alpha = .05$ and df = 4-1 =3, reject H_0 if $H \geq 7.81$.

Reject H_0. There is sufficient evidence to indicate that there is a difference in test scores for the four teaching techniques.

Key Concepts

I. Nonparametric Methods

1. These methods can be used when the data cannot be measured on a quantitative scale, or when
2. The numerical scale of measurement is arbitrarily set by the researcher, or when
3. The parametric assumptions such as normality or constant variance are seriously violated.

II. Wilcoxon Rank Sum Test: Independent Random Samples

1. Jointly rank the two samples: Designate the smaller sample as sample 1. Then

$$T_1 = \text{Rank of sample}_1 \quad \Rightarrow \quad T_1^* = n_1(n_1+n_2+1) - T_1$$

Key Concepts

2. Use T_1 to test for population 1 to the left of population 2

   Use T_1^* to test for population to the right of population 2.

   Use the smaller of T_1 and T_1^* to test for a difference in the locations of the two populations.
3. Table 7 of Appendix I has critical values for the rejection of H_0.
4. When the sample sizes are large, use the normal approximation: $z = \dfrac{T - \mu_T}{\sigma_T}$

$$\mu_T = \frac{n_1(n_1+n_2+1)}{2} \quad \text{and} \quad \sigma_T^2 = \frac{n_1 n_2(n_1+n_2+1)}{12}$$

Key Concepts

III. Sign Test for a Paired Experiment

1. Find x, the number of times that observation A exceeds observation B for a given pair.
2. To test for a difference in two populations, test $H_0 : p = 0.5$ versus a one- or two-tailed alternative.
3. Use Table 1 of Appendix I to calculate the p-value for the test.
4. When the sample sizes are large, use the normal approximation:

$$z = \frac{x - .5n}{.5\sqrt{n}}$$

Key Concepts

IV. Wilcoxon Signed-Rank Test: Paired Experiment

1. Calculate the differences in the paired observations. Rank the absolute values of the differences. Calculate the rank sums T^- and T^+ for the positive and negative differences, respectively. The test statistic T is the smaller of the two rank sums.
2. Table 8 of Appendix I has critical values for the rejection of for both one- and two-tailed tests.
3. When the sampling sizes are large, use the normal approximation:

$$z = \frac{T - [n(n+1)/4]}{\sqrt{[n(n+1)(2n+1)]/24}}$$

Key Concepts

V. Kruskal-Wallis *H* Test: Completely Randomized Design

1. Jointly rank the n observations in the k samples. Calculate the rank sums, T_i = rank sum of sample i, and the test statistic

$$H = \frac{12}{n(n+1)} \sum \frac{T_i^2}{n_i} - 3(n+1)$$

2. If the null hypothesis of equality of distributions is false, H will be unusually large, resulting in a one-tailed test.
3. For sample sizes of five or greater, the rejection region for H is based on the chi-square distribution with $(k - 1)$ degrees of freedom.

Key Concepts

VI. The Friedman F_r Test: Randomized Block Design

1. Rank the responses within each block from 1 to k. Calculate the rank sums $T_1, T_2, \ldots, T_k$, and the test statistic

$$F_r = \frac{12}{bk(k+1)} \sum T_i^2 - 3b(k+1)$$

2. If the null hypothesis of equality of treatment distributions is false, F_r will be unusually large, resulting in a one-tailed test.
3. For block sizes of five or greater, the rejection region for F_r is based on the chi-square distribution with $(k - 1)$ degrees of freedom.

Key Concepts

VII. Spearman's Rank Correlation Coefficient

1. Rank the responses for the two variables from smallest to largest.
2. Calculate the correlation coefficient for the ranked observations:

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad \text{or} \quad r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{if there are no ties}$$

3. Table 9 in Appendix I gives critical values for rank correlations significantly different from 0.
4. The rank correlation coefficient detects not only significant linear correlation but also any other monotonic relationship between the two variables.

Sensitivity vs. Specificity

- Sensitivity is a measure of the fraction of gold standard known examples that are correctly classified/identified.
- **Sensitivity= TP/(TP+FN)**
- Specificity is a measure of the fraction of negative examples that are correctly classified:
- **Specificity = TN/(TN+FP)**
- TP = True Positives
- FN = False Negatives
- TN = True Negatives
- FP = False Positives

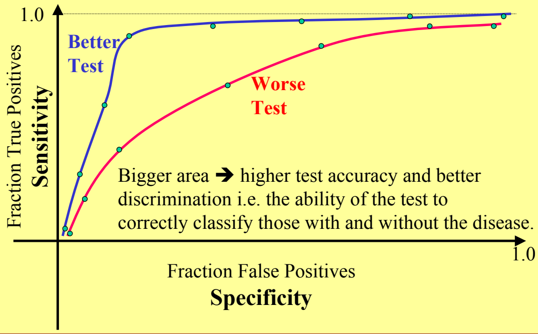
H_o: no effects (μ=0)

True Reality

Test Results	H_o true	H_o false
Can't reject	TN	FN
Reject H_o	FP	TP

The ROC Curve

- Receiver Operating Characteristic (ROC) curve



Receiver-Operating Characteristic curve

- ROC curve demonstrates several things:
  - It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
  - The closer the curve follows the left border and then the top border of the ROC space, the more accurate the test.
  - The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
  - The slope of the tangent line at a cut-point gives the likelihood ratio (LR) for that value of the test. You can check this out on the graph above.
  - The area under the curve is a measure of test accuracy.