The General Idea Behind Nonparametric Methods

/ μ, P, etc...

- All hypothesis tests and confidence intervals have thus far involved a population parameter. At times, we will be running a hypothesis test in which one doesn't exist. **Nonparametric** methods solve this problem.

- Our hypothesis tests have utilized a sampling distribution – usually the normal model, the Student's t model, or the Chi-Square model. Nonparametric methods are typically **distribution-free**.

Example: An absent-minded statistics professor would like to test if his day and night classes differ on final grades, but he lost the numerical scores of each of the students. He can print off from MyCecil the letter grades, though. Develop the ideas to test for a difference in class performance.

Conditions: Data values within each class must be independent. Check for randomness or an unbiased sample.

The groups must be independent of each other.

There is **no** requirement on sample size, no assumption about the distribution of the data or the errors, or no issues with outliers.

The cost for running nonparametric tests is a loss of statistical power.

To test the difference between two groups (we usually use a two-sample *t*-test), we use the Wilcoxon Rank-Sum / Mann-Whitney Statistic. (95% of the power)

Hypotheses: To test for a difference in class performance, we'd test these hypotheses:

H_0: There is no difference between night and day sections

H_A: There is a difference between night and day sections

The Data: For this example, we have the following final grades:

Day	**D**	**B+**	**C**	**A**	**A-**	**B**	**A**	**B**	
Night	**B+**	**C-**	**B**	**B-**	**D+**	**A-**	**F**	**B**	**C+**

We need to rank the values – it won't matter if we rank the lowest grade as "1" or the highest grade as "1", but we need to be consistent.

If there are ties, we average out the rank assigned to each value. Fill in the grades and ranks:

	D			D					D	D	D			D	D	D
F	D	D+	C-	C	C+	B-	B	B	B	B	B+	B+	A-	A-	A	A
1	2	3	4	5	6	7	9.5	9.5	9.5	9.5	12.5	12.5	14.5	14.5	16.5	16.5
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Thoughts: How to run this hypothesis test?

Clearly, if all the best grades (highest ranks) were in one class, we'd have a no-brainer decision. For example, if the best eight grades were all in the Day section:

$$T_{DAY} = 17 + 16 + \cdots + 10 = 108$$ (As big as it can get)

$$T_{NIGHT} = 1 + 2 + \cdots + 9 = 45$$ (As small as it can get)

If the grades alternated, then we'd have to conclude there wasn't much of a difference:

$$T_{DAY} = 2 + 4 + \cdots + 16 = 72$$

$$T_{NIGHT} = 1 + 3 + \cdots + 17 = 81$$

These are called rank sums

- To run the test, we don't need to compare rank sums. Rather, we can just look at the rank sum of one group – typically we pick the smaller group (with a fixed number of ranks, if we know one rank sum, we automatically know the other).

- For a sample size with eight values (the Day section), we could get as extremes:

Min: $T_{DAY} = 1 + 2 + \cdots + 8 = 36$

MAX: $T_{DAY} = 10 + 11 + \cdots 17 = 108$

- If the null hypothesis is true (no difference in sections), then the expected rank sum is right in the middle of these two extremes. That is, if there's no difference between groups, we'd expect the rank sum to be:

Expected Rank Sum | Ho True =

$$\frac{36+108}{2} = 72$$

- For the smaller group (Day Section), calculate the rank sum:

$$T_{DAY} = 2 + 5 + 9.5 + 9.5 + 12.5 + 14.5 + 16.5 + 16.5$$

$$= 86$$

On the high side, but is it high enough?

- If the null hypothesis is true, how unlikely is it to get a rank sum of 86 or higher? That probability is precisely the P-Value of the test.
- Usually we would shade under the appropriate sampling distribution model to determine the *P*-value, but we have no model. Rather, because there are only so many combinations of ranks and only so many possible values of the rank sum, we (actually others have) can consider all possible rankings!
- Under the null hypothesis, every possible ranking is equally likely, so running this test essentially boils down to counting the number of rank sums that exceed our rank sum. That proportion is exactly the *P*-value of the test.
- For tests with the smaller sample size is 10 or under, we can use the provided Table and get the exact *P*-value. Finish running the test using the table:

Test Stat is $T = 86$, running a two-tailed test.

Table gives $T_L = 54$ (lower bound)
$T_u = 90$ (upper bound)

for sample sizes of $n_1 = 8$ and $n_2 = 9$

The test is two-tailed so any T under 54 or over 91 is reject H_0 at $\alpha = 0.10$.

Since our T is not that extreme, P-Value would exceed 0.10, fail to reject H_0.

No evidence classes differ on grades.

Example: The urinary fluoride concentration (parts per million) was measured both for a sample of livestock grazing in an area previously exposed to fluoride pollution and for a similar sample grazing in an unpolluted area:

$n_1 = 5$
$n_2 = 7$

Polluted	21.3	~~18.7~~	23.0	~~20.1~~	~~16.8~~	~~20.9~~	~~19.7~~
Unpolluted	~~14.2~~	~~18.3~~	~~17.2~~	~~17.1~~	~~18.4~~		

Does the data strongly suggest that the true average fluoride concentration for livestock grazing in the polluted area is larger than for the unpolluted area? Run the Rank-Sum Test.

Rank	1	2	3	4	5	6	7	8	9	10	11	12
Group	U	P	U	U	U	U	P	P	P	P	P	P
Value	14.2	16.8	17.1	17.2	18.3	18.4	18.7	19.7	20.1	20.9	21.3	23.0

H_0: The Flouride Concentration is the same

H_A: The Polluted Group has a higher concentration

$T = 1 + 3 + 4 + 5 + 6 = 19$

Table provides $T_L = 20$ for one-tailed $\alpha = 0.025$ test.

Since $T = 19$ is under $T_L = 20$

Reject H_0 at $\alpha = 0.025$

There is evidence the Polluted Group has higher Flouride Concentration.

Normal Approximation for the Mann-Whitney Test

- A sum of even just 10 ranks has a distribution that is close enough to Normal for purposes of running the test for a difference in groups. To do the approximation, we need to find the mean and standard deviation.

- For groups 1 and 2, with corresponding rank sums of T_1 and T_2:

$$E(T_i) = n_i(n_1+n_2+1)/2 = \mu_i$$

$$SD(T_i) = \sqrt{n_1 n_2 (n_1+n_2+1)/12}$$

for $i = 1$ or 2

Example: In the following two samples, we have the urinary concentration of cotanine, a major metabolite of nicotine, from infants who had been exposed to household cigarette smoke and from a sample of unexposed infants.

Unexposed	8	11	12	14	20	43	111	
Exposed	35	56	83	92	128	150	176	208

Does the data suggest that the true average cotanine level is higher in exposed infants than in unexposed infants by more than 25? Carry out the Rank-Sum test using the Normal approximation at a 0.05 level of significance.

H_0: Exposed infants and unexposed infants have the same cotanine levels

H_A: Exposed infants have 25 more units of cotanine compared to non-exposed infants

Subtract 25 from each exposed value

	(1)	(3)	(4)	(5)	(6)	(8)	(12)	
UnExposed:	8	11	12	14	20	43	111	
Exposed:	10	31	58	67	103	125	151	183
	(2)	(7)	(9)	(10)	(11)	(13)	(14)	(15)

Get Ranks :

Un Exposed : $T_1 = 1+3+4+5+6+8+12$

$= 39$

Exposed : $T_2 = 2+7+9+10+11+13+14+15$

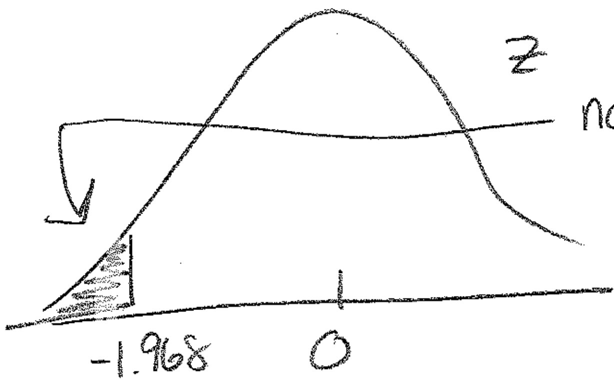
$= 81$

$$E[T_1] = 7(7+8+1)/2 = 56$$

$$SD[T_1] = \sqrt{7 \times 8(7+8+1)/12} \approx 8.64$$

$$Z = \frac{T_1 - E[T_1]}{SD[T_1]} = \frac{39-56}{8.64} = -1.968$$

Get P-Value



z ncdf(-E99, -1.968, 0, 1) $\approx$

0.025

At $\alpha = 0.05$, Reject H_0 and conclude the unexposed group has at least 25 fewer cotanine on average.

Example: A razor blade company claims that its "five-blade" disposable razor "gets you a lot more shaves" than any twin-blade razor on the market. Intrigued, a company than manufacturers twin-blade razors recruits 16 men and randomly assigns them to one of the two razors. The number of shaves that each gets before requesting a new razor is recorded below. Do the data support the five-blade claim at the 0.05 level of significance? Run the Normal approximation to the Rank-Sum test.

Ranks	3.5	7	8	9.5	11	12	15	16
Five-Blades	6	8	9	10	11	12	15	17
Twin-Blades	3	5	6	7	7	10	13	14
Ranks	1	2	3.5	5.5	5.5	9.5	13	14

H_0: Blades work Equally Well

H_A: Five Blades last longer than twin Blades.

$$T_{FIVE} = 3.5 + 7 + 8 + 9.5 + 11 + 12 + 15 + 16 = 82$$

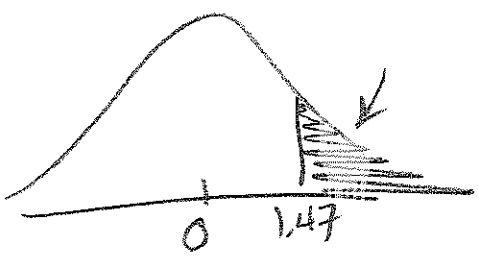
$$E\left[T_{FIVE}\right] = 8(8+8+1)/2 = 68$$

$$SD\left[T_{FIVE}\right] = \sqrt{8 \cdot 8(8+8+1)/12} \approx 9.52$$

$$Z = \frac{82-68}{9.52} = 1.47$$

Fail to Reject H_0

P-Value = ncdf(1.47, E99, 0, 1) = 0.071



At the $\alpha = 0.05$ level, there is no evidence the 5-Blade Razors last longer than the 2-Blade Razors.

The Kruskal-Wallace Test

- Just as the two-sample **t** test generalized to a one-way ANOVA, the Wilcoxon Rank-Sum test generalizes to the Kruskal-Wallace test when we are comparing the centers of three or more groups.
- The null hypothesis is that all groups have equal centers.
- The alternative hypothesis is that at least one group is different
- The test statistic is given by:

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

N = Total # obs, T_i are group rank sums, n_i are group sizes

- The test statistic approximately follows a chi-square model with *k* degrees of freedom (*k* is the number of groups). To find the *P*-value, always shade up and reject the null hypothesis if the *P*-value is small.
- For very small samples, tables exist for the Kruskal-Wallace test (but don't worry about it).

Example: The accompanying data contains the concentrations of the radioactive isotope strontium-90 in milk samples obtained from five randomly selected dairies in each of four different regions. Test at the 0.05 level to see whether the true average strontium-90 concentration differs in at least two regions.

Region	**1**	~~6.4~~	~~5.8~~	~~6.5~~	~~7.7~~	~~6.1~~
	2	~~7.1~~	~~9.9~~	~~11.2~~	~~10.5~~	~~8.8~~
	3	~~5.7~~	~~5.9~~	~~8.2~~	~~6.6~~	~~5.1~~
	4	~~9.5~~	12.1	~~10.3~~	12.4	~~11.7~~

H_0: Strontium-90 levels are equal in the 4 regions

H_A: At least one region differs

$n_1 = n_2 = n_3 = n_4 = 5$

$N = 20$

Rank	Value	Group
1	5.1	3
2	5.7	3
3	5.8	1
4	5.9	3
5	6.1	1
6	6.4	1
7	6.5	1
8	6.6	3
9	7.1	2
10	7.7	1
11	8.2	3
12	8.8	2
13	9.5	4
14	9.9	2
15	10.3	4
16	10.5	2
17	11.2	2
18	11.7	4
19	12.1	4
20	12.4	4

$T_1 = 3+5+6+7+10 = 31$

$T_2 = 9+12+14+16+17 = 68$

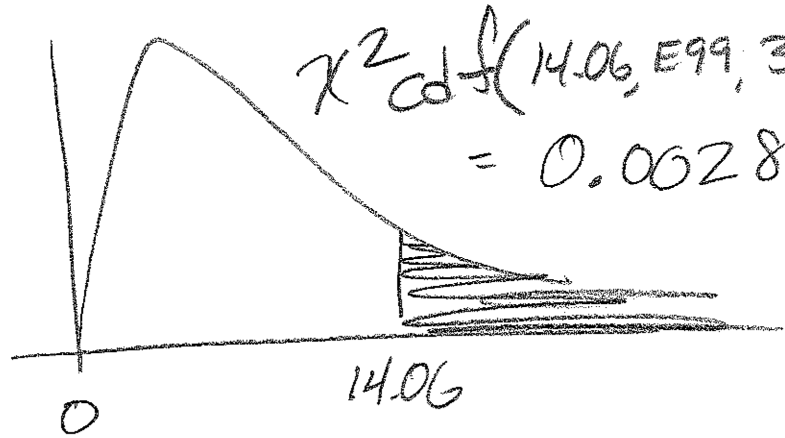
$T_3 = 1+2+4+8+11 = 26$

$T_4 = 13+15+18+19+20 = 85$

$$H = \frac{12}{20(21)}\left(\frac{31^2}{5} + \frac{68^2}{5} + \frac{26^2}{5} + \frac{85^2}{5}\right) - 3(20+1)$$

$= 14.06$

$H \sim \chi^2$ df=3



χ^2cdf(14.06, E99, 3)

$= 0.0028$

Reject Ho, conclude there is a diffrence in strontium-90 concentrations at the 4 dairies.

Example: A drug company tested three dosages of a pain medicine by randomly recruiting 27 people with chronic headaches. Nine people each were randomly assigned to drug A, B, or C and reported during their next headache 1 = "no pain" up to 10 = "extreme pain".

Use the Kruskal-Wallace test statistic to determine if there is a difference among the drugs.

Drug	Pain	Drug	Pain	Drug	Pain
A	4	B	6	C	6
A	5	B	8	C	7
A	4	B	4	C	6
A	3	B	5	C	6
A	2	B	4	C	7
A	4	B	6	C	5
A	3	B	5	C	6
A	4	B	8	C	5
A	4	B	6	C	5

Rank	Pain	Drug
1	2	A
2.5	3	A
2.5	3	A
7	4	A
7	4	A
7	4	A
7	4	A
7	4	A
7	4	B
7	4	B
13.5	5	A
13.5	5	B
13.5	5	B
13.5	5	C
13.5	5	C
13.5	5	C

Rank	Pain	Drug
20	6	B
20	6	B
20	6	B
20	6	C
20	6	C
20	6	C
20	6	C
24.5	7	C
24.5	7	C
26.5	8	B
26.5	8	B

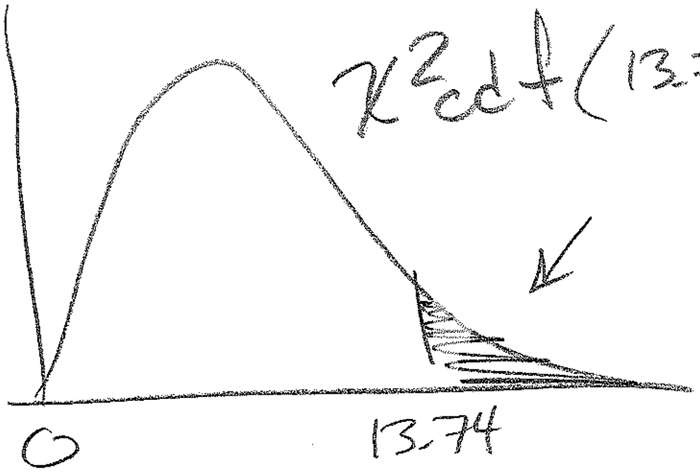
$T_A = 54.5$ $T_B = 154$

193 ~~192~~ $T_C = 169.5$

$n_1 = n_2 = n_3 = 9$ $N = 27$

$$H = \frac{12}{27(28)}\left(\frac{54.5^2}{9} + \frac{154^2}{9} + \frac{169.5^2}{9}\right) - 3(27+1)$$

$H = 13.74$ with $k=3$



$\chi^2cdf(13.74, E99, 3) = 0.003$

Reject H_0.

There is strong evidence that the effectiveness of Drugs A, B, C differs for alleviating headaches.