Testing Paired Data

- Previously, we tested two dependent samples (paired data) with a *t*-test. We would take differences for each pair of data and, presuming the sample size was large or the differences were plausibly Normal, we would proceed with the parametric test.
- The nonparametric equivalent is called the Wilcoxon Signed-Rank Test
- Step 1: Find the pairwise difference for each subject.
- Step 2: Omit any differences of zero.
- Step 3: Rank the **absolute value** of the differences.
- The null hypothesis is that:

H_0: The Differences are Centered at zero (no difference in treatments)

- Step 4: Determine the test statistic. There will be two rank sums:

T^- is the sum of the ranks with neg. differences

T^+ is the sum of the ranks with pos. differences

Using only the smaller sum use the Table of Critical Values to run the test.

- Assumptions and Conditions: The data must be paired and the pairs must be mutually independent.

Critical Values of the Wilcoxon Signed Ranks Test

n	Two-Tailed Test		One-Tailed Test	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

Example: Trace metals in drinking water affect the flavor, and unusually high levels can pose a health risk. Reports on a study in which six river locations were selected and analyzed for zinc concentration (mg/L). Measurements in the six locations were taken at the surface and at the bottom of the river. Does the data suggest that the true median concentration in the bottom water exceeds that at the surface? Run the Wilcoxon Signed-Rank Test.

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Zinc at Bottom	0.430	0.266	0.567	0.531	0.707	0.716
Zinc at Surface	0.440	0.238	0.390	0.410	0.605	0.755
Difference	−0.01	0.028	0.177	0.121	0.102	−0.039
Absolute Value Diff.	0.01	0.028	0.177	0.121	0.102	0.039
Rank	1	2	6	5	4	3

H_0: There is no difference in zinc levels at the surface versus the bottom

H_A: The bottom has more zinc.

$T^- = 1 + 3 = 4$

$T^+ = 2 + 4 + 5 + 6 = 17$

Test Stat is T^-

Table: $n = 6$

One-tailed $\alpha = 0.05$ test needs T^- to be under 2.

Fail to Reject. No evidence to say bottom has more zinc.

$n = 6$, hard to run t test on differences without believing they're normal

Example: A recent sample of airfares from New York City to various locations across the country was conducted to see if airfare has increased over the past year. Run the Wilcoxon Signed-Rank test to determine if it has.

City	This Year	Last Year	Difference	Absolute Value of the Difference	Rank
Cincinnati	575.67	511.11	64.56	64.56	10
Dallas	198.98	185.12	13.86	13.86	5
Boston	220.20	225.25	-5.05	5.05	2
Miami	399.66	340.25	59.41	59.41	9
Las Vegas	341.05	335.95	5.10	5.10	3
Los Angeles	410.50	324.93	85.57	85.57	11
Seattle	487.87	478.21	9.66	9.66	4
Detroit	230.25	249.49	-19.24	19.24	7
Memphis	298.15	296.53	1.62	1.62	1
Columbus	336.90	320.19	16.71	16.71	6
Denver	426.88	401.49	25.39	25.39	8
Portland	477.65	~~425.57~~ 477.65	0	X	X

$n = 11$

$T^- = 2 + 7 = 9$ ← Test Stat

$T^+ = 10 + 5 + 9 + 3 + 11 + 4 + 1 + 6 + 8 = 57$

H_0: This Year & Last Year have same median Airfare

H_A: This year's median is higher

Reject at $\alpha = 0.05$ (Cut Point is 13)

Fail to Reject at $\alpha = 0.01$ (Cut Point is 7)

There is some evidence Prices are increasing.

When Should We Use a Nonparametric Method?

- If your data contain information only about order, then you're basically stuck with a nonparametric method.
- If your quantitative data violates one of the conditions or assumptions (small sample size, outliers, non-Normal), nonparametric methods can be of value.
- Translating quantitative data to ranks essentially hides the outliers, protects us from data with two or more modes, and ignores problems associated with analyzing skewed datasets.
- The "cost" of this service is a loss of statistical power.
- Don't default to nonparametric methods. If the conditions are met, run t-tests and ANOVA because these methods are generally better and more powerful.
- There are nonparametric methods for correlation and regression as well, we just didn't cover them in this class.
- Always ensure your data values are independent of each other. The P-values of these nonparametric tests is based on every rank being equally likely. For that to be true, we need independence.

Example: Midterm Scores
(Compare Two Versions with a Randomization Test)

- Back in Spring 2011, a Math 127 professor gave two versions of the midterm to the twenty students – here are the data.
- As occasionally happens, the students began to complain that the Green Version was harder and therefore "unfair" to those who took that test.

Green	70	78	62	93	51	72	58	84	91	62
Yellow	60	53	88	75	96	85	68	90	90	76

Notes: $\bar{y}_{green} = 72.1$

$\bar{y}_{yellow} = 78.1$

$D = -6$ points

Activity (work with one other person):

1. Take the 20 cards (each has a test score) and shuffle them up very thoroughly. Deal out the cards into two piles (at random, of course). List the scores below.

Green										
Yellow										

2. Calculate the mean score for each group:

Mean score for the Green Version: _____

Mean score for the Yellow Version: _____

Subtract Green – Yellow: _____ (sign matters, plus or minus!)

3. Write your mean difference on the post-it note and place it on the white board.

4. **The main question** – "Was our difference of 6 points between the midterm versions because the Green version was harder or was the difference due to the Randomization of test versions as students walked into the room?"

We have competing theories here:

Theory A (Skeptical Theory): Test Versions Equally Difficult

Theory B (Alternative Explanation Theory): Green Version is Harder

5. Now look at the post-it notes on the board.

The biggest difference in mean scores was _____.

The number of differences that were at least 6 points was _____.

The proportion of differences that were at least 6 points was _____.

Are we convinced that **Theory A** or **Theory B** is the best choice? What should we do?

Collect More Data.

- Post it notes are inefficient
- StatCrunch Simulation

6. The instructor will run a simulation on StatCrunch. Altogether, we will shuffle the deck and calculate the mean difference 3007 times.

7. Using the simulation data, the proportion of differences that were at least 6 points was ~0.263. We call this a P-Value.

P-Value Definition (for this example):

If the versions of the midterm are no different (skeptical theory), the P-Value is the proportion of post-it notes that would have mean differences of at least 6 points when we randomize test versions.

When we randomized 3007 times, we got a P-Value of 0.363.

Is this proportion unusual? Not at all.

- Said another way, **if the versions of the test are no different**, we would get a mean difference of at least 6 points 36.3 % of the time if we randomize the midterm version scores.

8. Which competing theory should we go with?

Theory A (Versions of the midterm are no different)

Theory B (The Green Version is harder)

The 6 point difference could have happened by chance

9. Did we just prove conclusively that the versions are equal in difficulty? No!

What did we show?

10. Statistical Significance:

If your results are surprising, assuming the skeptical theory is true, we say the data are statistically significant.

For this example, the actual midterm scores are / are not statistically significant.

The professor should / should not feel satisfied that the versions were probably fair.