Alternatives to t-Tools

PAIRED T-TEST

NONPARAMETRIC PAIRED TESTS

Paired or Matched t-Tools

Paired t-Test

Known alternatively as Matched Pairs or Dependent t-Test

Assumptions

- Data are either:
  - From one sample that has been test twice (example pre and post test / repeated measures)
  - Or from a group of subjects that are thought to be similar and can thus be matched or paired (example from same family, or twins)
- Data are Normally distributed, independent between observations, and have identical variances.

Example of repeated measures

Number	Name	Test 1	Test 2
1	Mike	35%	67%
2	Melanie	50%	46%
3	Melissa	90%	86%
4	Mitchell	78%	91%

Example of matched pairs

Pair	Name	Age	Test
1	John	35	250
1	Jane	36	340
2	Jimmy	22	460
2	Jessy	21	200

A Look at the Variance

- Suppose Y_{before} and Y_{after} are the before and after measurements for one subject

- Fact: $Variance(Y_{before} - Y_{after}) = \sigma^2_{before} + \sigma^2_{after} - 2\, Covariance(Y_{before}, Y_{after})$

- Due to the assumption about equal variances for before/after:

$$Variance(Y_{before} - Y_{after}) = 2\,\sigma^2 - 2\, Covariance(Y_{before}, Y_{after})$$

- This is different than the two-sample t-test:

$$\text{Variance}(\bar{Y}_{before} - \bar{Y}_{after}) = \frac{\sigma^2_{before}}{n_{before}} + \frac{\sigma^2_{after}}{n_{after}}$$

Example:
Medical Reasoning Test

- The AMA has a diagnostic test for medical reasoning
- On average, people score about 500 points on this test
- We have data from 10 subjects who took the medical reasoning test. These subjects were randomly selected from St. Paul Hospital in Dallas

- **Not fatigued:** is the baseline, taking the test before a shift
- **Fatigued:** is after the treatment; working for 12 operational hours prior to re-taking the test

Subject #	Not Fatigued	Fatigued
1	567	530
2	512	492
3	509	510
4	593	580
5	588	600
6	491	483
7	520	512
8	588	575
9	529	530
10	508	490

(Lower numbers = worse score)

Example:
Medical Reasoning Test

A scientific question would be if there is a decrease in medical reasoning after a long shift

To convert this to a statistical question, we can try to test whether the DIFFERENCE OF THE MEANS between the fatigued scores and the non fatigued scores is less than zero

$$H_0: \mu_{fatigued} - \mu_{not\ fatigued} = 0$$

$$H_A: \mu_{fatigued} - \mu_{not\ fatigued} < 0$$

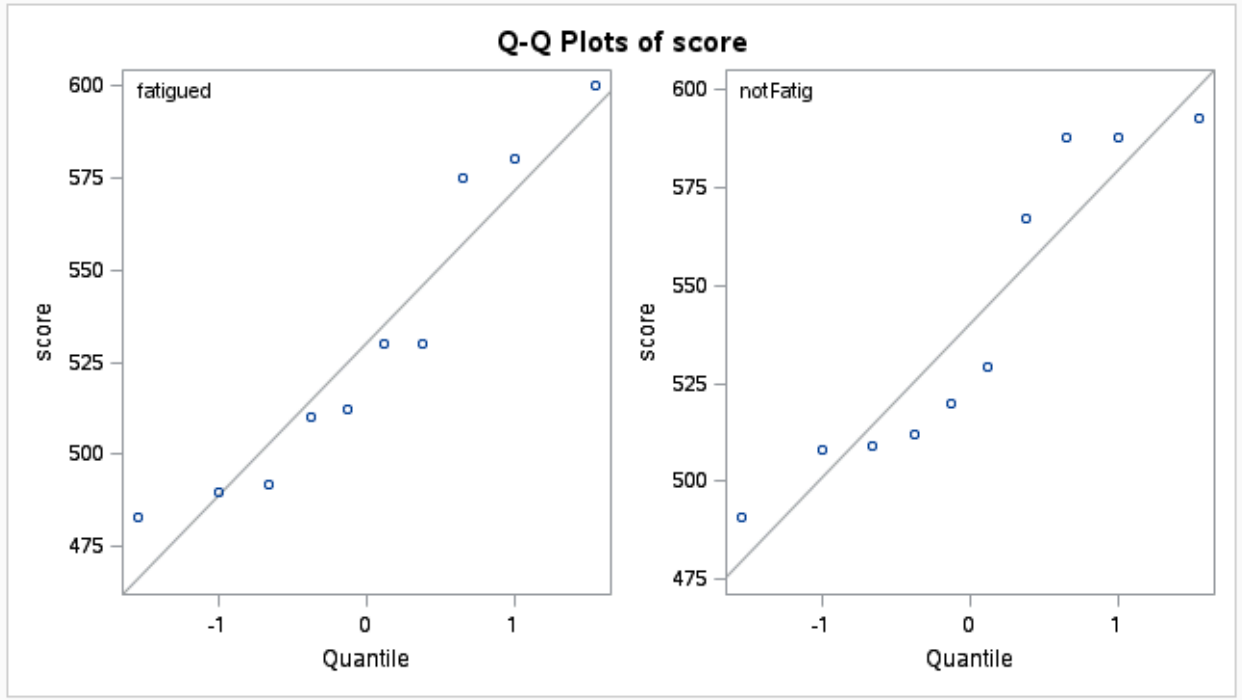
Since this is a medical application, let's use a 0.01 decision rule for the p-value

Example: Medical Reasoning Test

If we did this, we would be wrong! Why?

A fundamental assumption is violated: independence

```
PROC TTEST DATA=mrt ALPHA = 0.01 SIDE = L;
    CLASS status;
    VAR score;
RUN;
```



The TTEST Procedure
Variable: score

status	N	Mean	Std Dev	Std Err	Minimum	Maximum
fatigued	10	530.2	41.3677	13.0816	483.0	600.0
notFatig	10	540.5	39.2067	12.3983	491.0	593.0
Diff (1-2)		-10.3000	40.3017	18.0235		

status	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
fatigued		530.2	487.7	572.7	41.3677	25.5520	94.2197
notFatig		540.5	500.2	580.8	39.2067	24.2172	89.2977
Diff (1-2)	Pooled	-10.3000	-Infty	35.7027	40.3017	28.0506	68.3134
Diff (1-2)	Satterthwaite	-10.3000	-Infty	35.7155			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	18	-0.57	0.2874
Satterthwaite	Unequal	17.948	-0.57	0.2874

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	1.11	0.8756

Assumption Check Failure

We need to account for the dependence between the two groups

A Look (back) at the Variance

- The core result for the two-sample t-test depends crucially on independence

$$\text{Variance}(\bar{Y}_{not\ fatigued} - \bar{Y}_{fatigued}) = \frac{\sigma^2_{not\ fatigued}}{n_{not\ fatigued}} + \frac{\sigma^2_{fatigued}}{n_{fatigued}}$$

- As the subjects are the same people, the scores are very dependent

- In fact, the scores are usually positively correlated ($Covariance\left(Y_{not\ fatigued}, Y_{fatigued}\right) > 0$)

$$\text{Variance}\left(Y_{not\ fatigued} - Y_{fatigued}\right) = \sigma^2_{not\ fatigued} + \sigma^2_{fatigued} - 2\ Cov\left(Y_{not\ fatigued}, Y_{fatigued}\right)$$

- Therefore, the actual variance is smaller than if the samples were independent

Example:
Medical Reasoning Test

Instead of testing the DIFFERENCE OF THE MEANS:

$$H_A: \mu_{fatigued} - \mu_{not\ fatigued} < 0$$

We should test the MEAN OF THE DIFFERENCES:

$$H_o: \mu_{fatigued-not\ fatigued} = 0$$
$$H_A: \mu_{fatigued-not\ fatigued} < 0$$

Subject	Not Fatigued	Fatigued	Difference
1	567	530	-37
2	512	492	-20
3	509	510	1
4	593	580	-13
5	588	600	12
6	491	483	-8
7	520	512	-8
8	588	575	-13
9	529	530	1
10	508	490	-18

Paired T-test

Subject	Not Fatigued	Fatigued	Difference (d_i)
1	567	530	-37
2	512	492	-20
3	509	510	1
4	593	580	-13
5	588	600	12
6	491	483	-8
7	520	512	-8
8	588	575	-13
9	529	530	1
10	508	490	-18

$$\bar{d} = \frac{d_1 + d_2 + \ \ldots + d_{10}}{10}$$

s_d is the sample std. dev.

$$SE(\bar{d}) = \frac{s_d}{\sqrt{10}}$$

$$T = \frac{\bar{d}}{SE(\bar{d})}$$

A SAS Code Comparison

```
DATA mrt;
    INPUT score status $ @@;
    DATALINES;
567 notFatig 512 notFatig 509 notFatig 593
491 notFatig 520 notFatig 588 notFatig 529
530 fatigued 492 fatigued 510 fatigued 580
483 fatigued 512 fatigued 575 fatigued 530
;
RUN;
```

```
PROC TTEST DATA=mrt ALPHA = 0.01 SIDE = L;
    CLASS status;
    VAR score;
RUN;
```

(Two sample T-Test)

```
DATA mrt_paired;
    INPUT fatigued notFatig @@;
    DATALINES;
567 530 512 492 509 510 593 580 588 600
491 483 520 512 588 575 529 530 508 490
;
RUN;
```

```
PROC TTEST DATA=mrt_paired ALPHA = 0.01 SIDE = L;
    PAIRED fatigued*notFatig;
RUN;
```

(Paired T-test)

A SAS Code Comparison

The TTEST Procedure
Variable: score

status	N	Mean	Std Dev	Std Err	Minimum	Maximum
fatigued	10	530.2	41.3677	13.0816	483.0	600.0
notFatig	10	540.5	39.2067	12.3983	491.0	593.0
Diff (1-2)		-10.3000	40.3017	18.0235		

status	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
fatigued		530.2	487.7	572.7	41.3677	25.5520	94.2197
notFatig		540.5	500.2	580.8	39.2067	24.2172	89.2977
Diff (1-2)	Pooled	-10.3000	-Infty	35.7027	40.3017	28.0506	68.3134
Diff (1-2)	Satterthwaite	-10.3000	-Infty	35.7155			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	18	-0.57	0.2874
Satterthwaite	Unequal	17.948	-0.57	0.2874

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	1.11	0.8756

(Two sample T-Test)

The TTEST Procedure
Difference: fatigued - notFatig

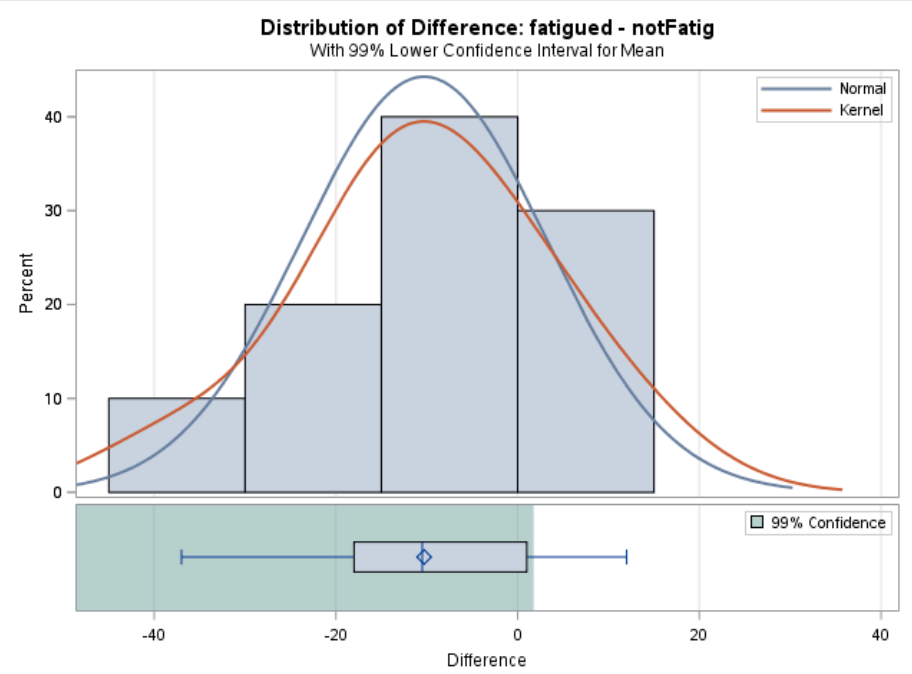
N	Mean	Std Dev	Std Err	Minimum	Maximum
10	-10.3000	13.5158	4.2741	-37.0000	12.0000

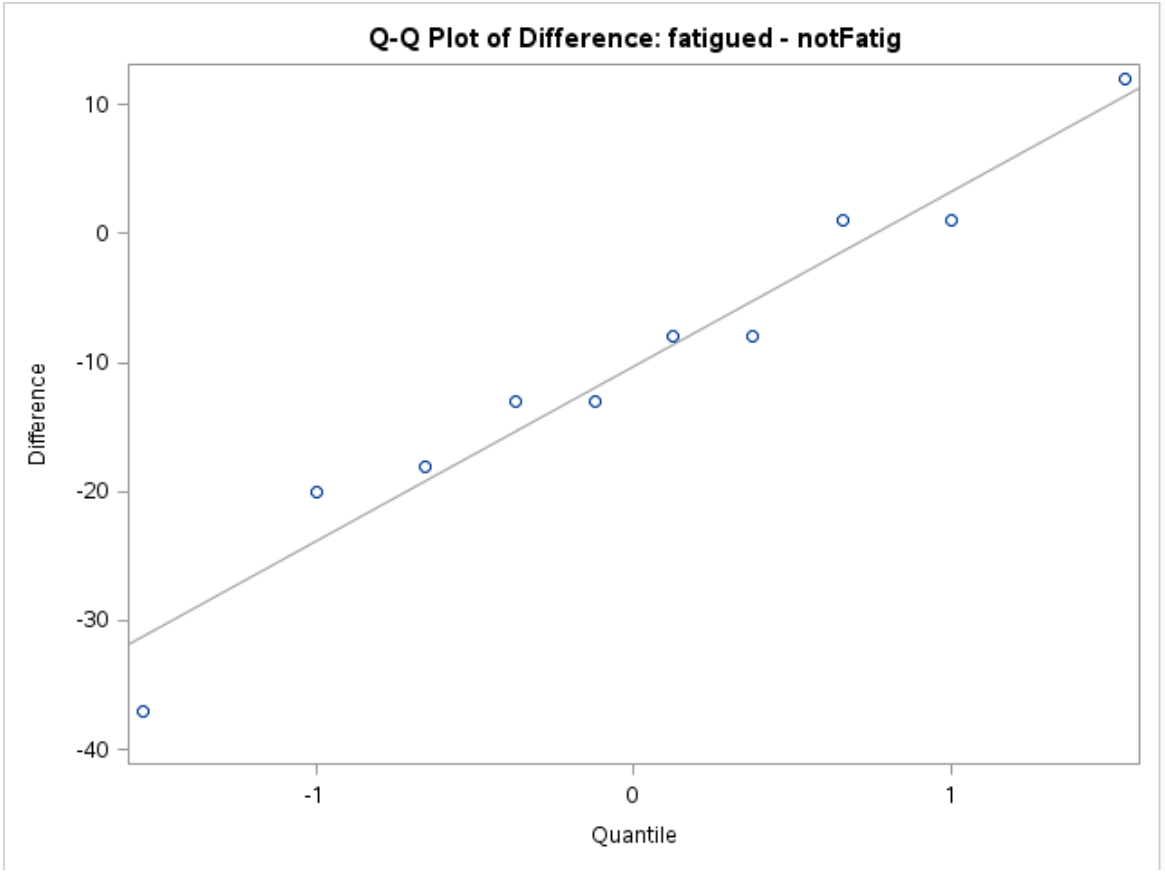
Mean	99% CL Mean		Std Dev	99% CL Std Dev	
-10.3000	-Infty	1.7591	13.5158	8.3485	30.7838

DF	t Value	Pr < t
9	-2.41	0.0196

(Paired T-test)

Checking the Assumptions

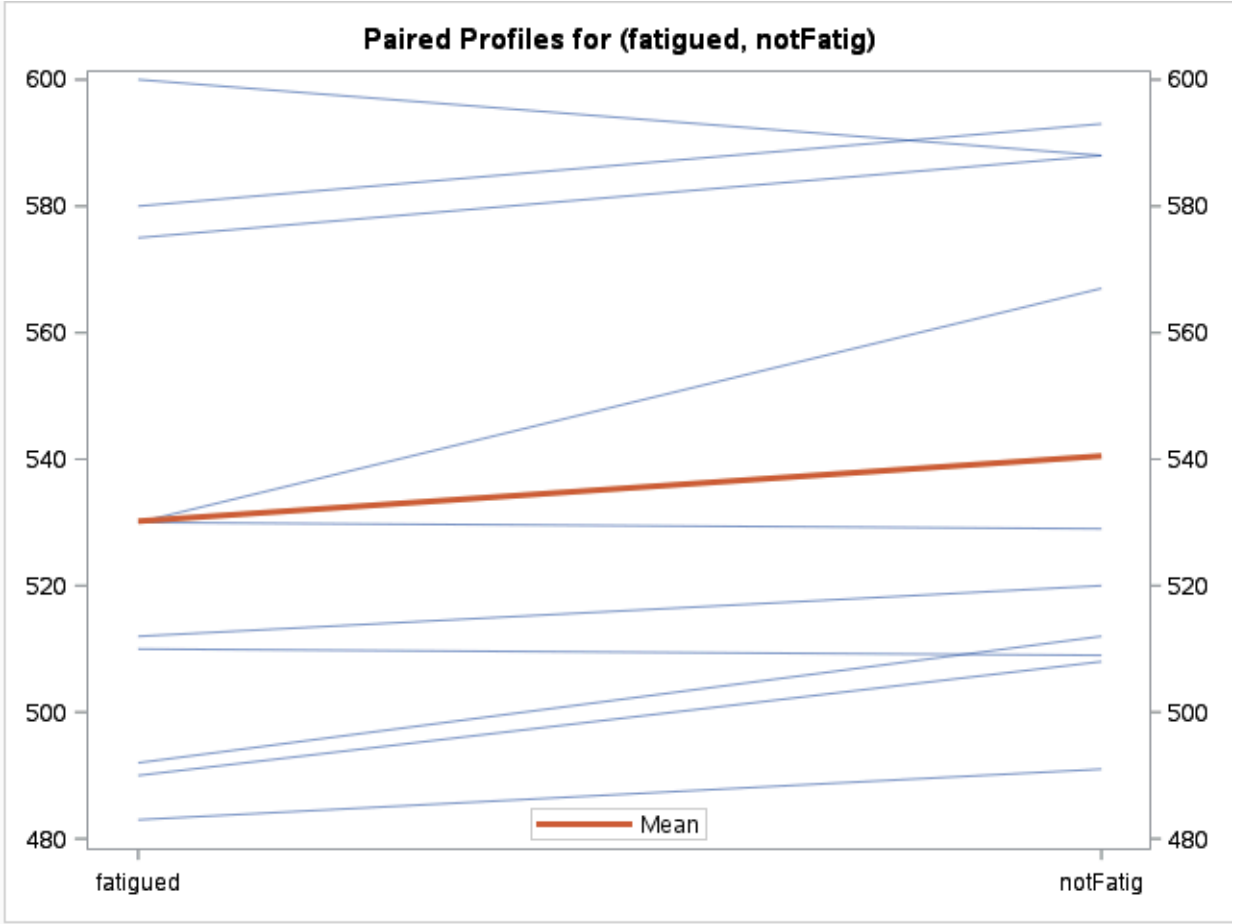




Also: is it reasonable for all the differences to be independent with the same variance?

Additional Information

- We can look at a **PROFILE PLOT**
- The lines connect the scores on the MRT in the “fatigued” versus “not fatigued” states for each subject



Conclusion

Statistical Conclusion: While there is some evidence that, on average, the fatigued subjects score lower than the non-fatigued subjects, we find the evidence not compelling enough for this application (paired one-sided t-test p-value = .0196) and hence conclude that there is no mean difference in MRT results. A 99% one tailed confidence interval for the mean difference in scores finds that a maximum plausible value for the mean difference is 1.76 points.

Scope of Inference: Since this was a random sample from St. Paul Hospital in Dallas, we can extend this result to the entire hospital.

Question: Would this maximum plausible value be positive or negative for a 95% one-tailed confidence interval?

Answer: Negative

Alternatives to the t-Test for Paired Data

Example: Nerve Site Data

```
/* Sign Test and Signed Rank Test */

data horse;
 input horse       site1       site2;
datalines;
6        14.2        16.4
4        17        19
8        37.4        37.6
5        11.2        6.6
7        24.2        14.4
9        35.2        24.4
3        35.2        23.2
1        50.6        38
2        39.2        18.6
;
```

For each of the 9 horses, a veterinary anatomist measured the density of nerve cells at specified sites in the intestine.

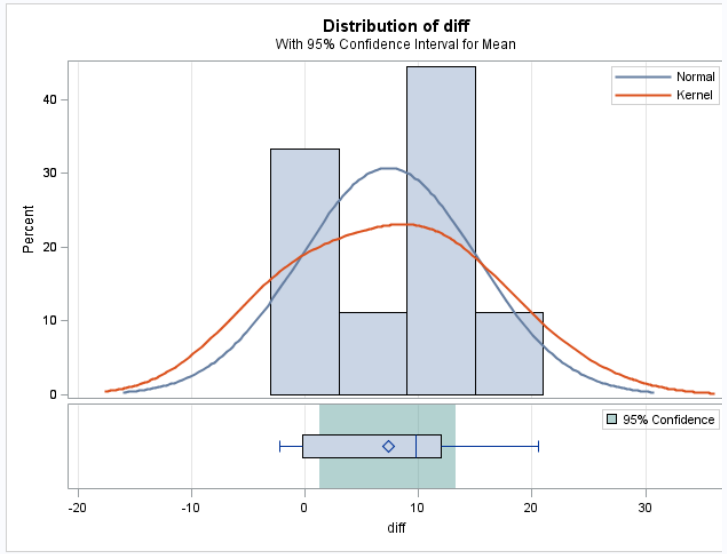
horse	site1	site2
6	14.2	16.4
4	17	19
8	37.4	37.6
5	11.2	6.6
7	24.2	14.4
9	35.2	24.4
3	35.2	23.2
1	50.6	38
2	39.2	18.6

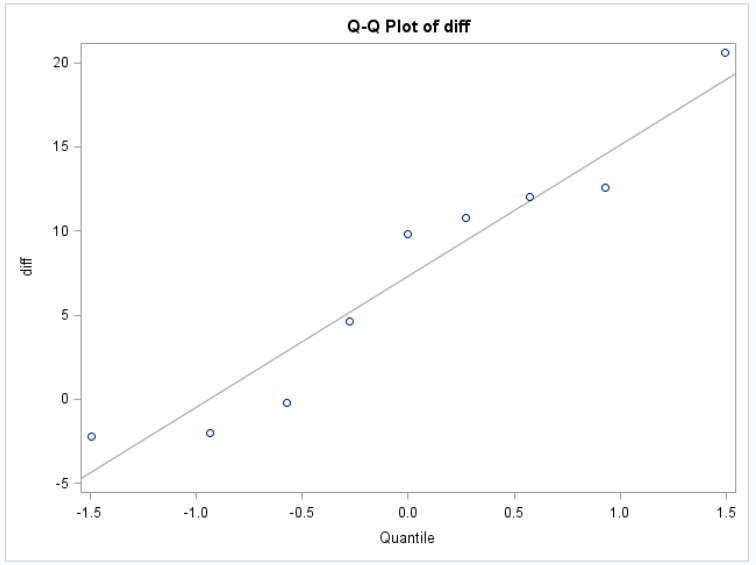
Using the paired t-Test

N	Mean	Std Dev	Std Err	Minimum	Maximum
9	7.3333	7.7929	2.5976	-2.2000	20.6000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
7.3333	1.3431	13.3235	7.7929	5.2638	14.9295

DF	t Value	Pr > |t|
8	2.82	0.0224





The sample size is rather small, hence the normality assumption is suspect

The Sign Test: Discussion and Assumptions

- No distributional assumptions and resistant to outliers
- It still requires some assumptions:
  1. Differences between paired observations are independent of other differences and come from the same distribution
  2. The *Y* values should be **INTERVAL** (though sometimes it is used with **ORDINAL** data as well)
- Can also be used as a 1-sample test for medians (instead of means)

The Sign Test: Hypotheses

The hypotheses will be in terms of **MEDIANS** instead of means

The Null Hypothesis:

H_0: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is zero

(Equivalently, H_0: Probability("site 1" > "site 2") = 0.5)

The Alternative Hypotheses:

H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is not zero **(TWO SIDED)**

(Equivalently, H_A: Probability("site 1" > "site 2") $\neq$ 0.5)

H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is greater than zero **(ONE SIDED)**

(Equivalently, H_A: Probability("site 1" > "site 2") $>$ 0.5)

The Sign Test: Nerve Site Data

H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is greater than zero

Under the alternative, we would expect that the # of positive differences, "site 1" - "site 2" > 0, should be large (call this number K)

(has a standard normal reference dist. under H_0. We can also use the binomial dist. to compute exact p-value)

$$z = \frac{K - 0.5 - {}^{n}/_{2}}{\sqrt{{}^{n}/_{4}}}$$

$$= \frac{6 - 0.5 - {}^{9}/_{2}}{\sqrt{{}^{9}/_{4}}} = 2/3$$

$$P(Z > 2/3) = 0.2527$$

(ONE SIDED, CC P-VALUE)

horse	site1	site2	diff	Sign
8	37.4	37.6	-0.2	-
4	17	19	-2	-
6	14.2	16.4	-2.2	-
5	11.2	6.6	4.6	+
7	24.2	14.4	9.8	+
9	35.2	24.4	10.8	+
3	35.2	23.2	12	+
1	50.6	38	12.6	+
2	39.2	18.6	20.6	+

K = 6

The Signed-Rank Test: Discussion and Assumptions

Nearly the same as for the sign test, save for the **bolded comments** below:

- **Almost** no distributional assumptions and resistant to outliers
- It still requires some assumptions:
  1. Differences between paired observations are independent of other differences and come from the same distribution
  2. The *Y* values should be **INTERVAL** (though sometimes it is used with **ORDINAL** data as well)
  3. **The distribution of the differences is symmetric (see next slide for details)**
- Can also be used as a 1-sample test for medians (instead of means)

The Signed-Rank Test: Hypotheses

The Signed-Rank test technically measures something known as the **PSEUDO-MEDIAN**

(in this case, the "pseudo-median" is the median of all pairwise differences of the differences)

If the distribution of the differences is symmetric, then the hypotheses will be in terms of **MEDIANS** (we will make this assumption in what follows to cut down on technicalities)

In this case, the null/alternative hypotheses are the same as for the sign test:

The Null Hypothesis:

H_0: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is zero

The Alternative Hypotheses:

H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is not zero **(TWO SIDED)**

H_A: The **MEDIAN** difference in nerve cell count between "site 1" and "site 2" is greater than zero **(ONE SIDED)**

$$\text{Mean}(S) = n(n+1)/4 \quad \text{and} \quad \text{SD}(S) = [n(n+1)(2n+1)/24]^{1/2}.$$

The Signed-Rank Test

Under the alternative, we would expect that the # of positive differences, K, should be large **AND** the sum of the ranks of the magnitudes for those K differences, |"site 1" - "site 2"|, should large as well (call this number S)

$$z = \frac{S - 0.5 - \text{Mean}(S)}{\text{SD}(S)}$$

$$= \frac{39 - 0.5 - (9 * 10)/4}{\sqrt{9 * 10 * 19/24}} = 1.896$$

(has a standard normal reference dist. under H_0)

$$P(Z > 1.896) = 0.029$$

(ONE SIDED, CC P-VALUE)

horse	site1	site2	abs(diff)	Sign	rank
8	37.4	37.6	0.2	-	1
4	17	19	2	-	2
6	14.2	16.4	2.2	-	3
5	11.2	6.6	4.6	+	4
7	24.2	14.4	9.8	+	5
9	35.2	24.4	10.8	+	6
3	35.2	23.2	12	+	7
1	50.6	38	12.6	+	8
2	39.2	18.6	20.6	+	9

S = 39

These are two sided p-values, so the one-sided p-values are half as large

Signed and Signed-Rank tests: Horse Data

```
/* Sign Test and Signed Rank Test */

data horse;
 input horse       site1       site2;
datalines;
6        14.2        16.4
4        17        19
8        37.4        37.6
5        11.2        6.6
7        24.2        14.4
9        35.2        24.4
3        35.2        23.2
1        50.6        38
2        39.2        18.6
;
```

```
data horse2;
set horse;
diff = site1 - site2;
run;

proc univariate data = horse2;
var diff;
run;
```

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2.823066	Pr > |t|	0.0224
Sign	M	1.5	Pr >= |M|	0.5078
Signed Rank	S	16.5	Pr >= |S|	0.0547

Conclusion and Some Notes

Statistical Conclusion: There is strong evidence that the distribution of nerve density at site 1 is larger than the distribution of nerve density at site 2 (Wilcoxon signed-rank test one-sided p-value of 0.0274). This means that for any value for nerve density, the probability of getting that value at site 1 is larger than the probability of getting that value at site 2.

Note:

- The signed-rank test has more power than the sign test

  (Compare the p-values 0.254 vs. 0.0274)

- Both tests make very few assumptions about the distributions