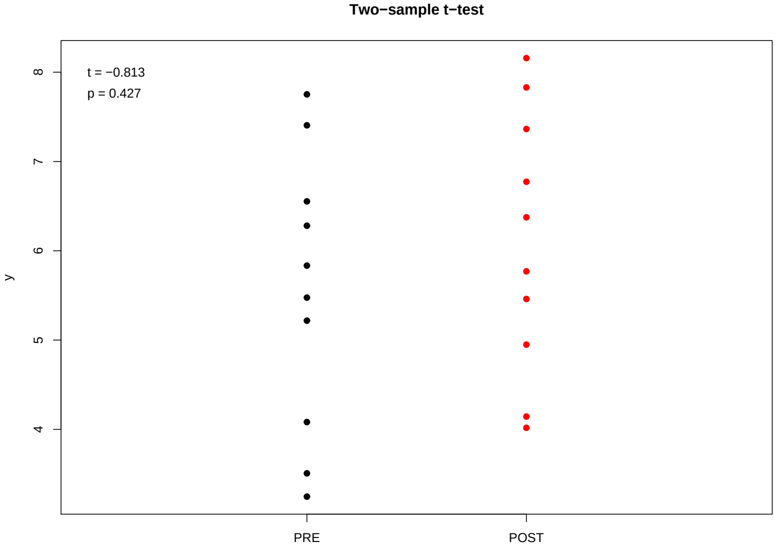
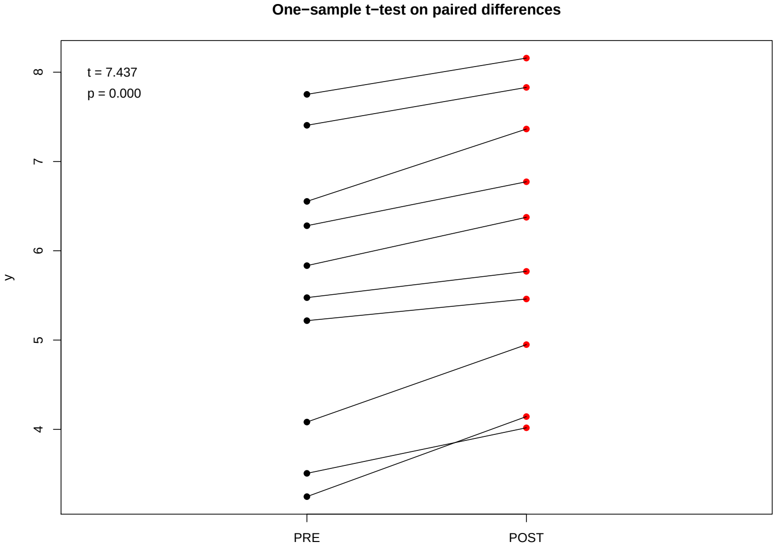
The Wilcoxon Signed Rank Test

(and why we drop the zeros)

Without Pairing: No rejection of the null.



With Pairing: Rejection of the null.



The Wilcoxon Signed Rank Procedure

For paired pre-treatment $(y_{11}, \ldots, y_{n1})$ and post-treatment: $(y_{21}, \ldots, y_{n2})$ data, the procedure is as follows:

1. Form $x_i = y_{i1} - y_{i2}$, for $i = 1, \ldots, n$.
2. Form $s_i = |x_i|$, for $i = 1, \ldots, n$.
3. Drop every $x_i = 0$, leaving a sample size of m.
4. Assign ranks $r_1, \ldots, r_m$ to $s_1, \ldots, s_m$ after sorting into ascending order.
5. Form

$$T_+ = \sum_{i=1}^{m} r_i z_i \qquad\qquad T_- = \sum_{i=1}^{m} r_i (1 - z_i)$$

where

$$z_i = \begin{cases} 1 & x_i > 0 \\ 0 & x_i < 0 \end{cases}$$

Computing the Null Distribution

The behaviour of T_+ and T_- can be predicted under the null hypothesis under the assumption that the original data are generated from **continuous distributions** with the same **median**.

- Under this assumption, we will **never** get $x_i = 0$.
- Under this assumption, if K is the random variable recording the **number** of positive x_i values, then

  $$K \sim Binomial(n, 1/2)$$

  as in the **sign test**.
- Under this assumption, given that $K = k$, T_+ is the sum of k numbers (ranks) chosen at random with replacement from the set $\{1, 2, \ldots, n\}$. Similarly, T_- is the sum of the remaining $n - k$ ranks.

By enumerating all the possible selections, we can compute the probability distribution (conditional on $K = k$) of T_+; denote it

$$p(t|k).$$

Then, using the **Theorem of Total Probability** and the binomial distribution formula, we can obtain the null distribution as

$$\begin{aligned} P[T_+ = t] &= \sum_{k=0}^{n} P[T_+ = t|K = k]\, P[K = k] \\ &= \sum_{k=0}^{n} p(t|k) \binom{n}{k} \left(\frac{1}{2}\right)^n \end{aligned}$$

An equivalent calculation holds for T_-; in fact, under H_0, the distributions of T_+ and T_- are identical.

Example: $n = 5$

We have to enumerate all possible sums of k numbers chosen from $\{1, \ldots, n\}$, for each $k = 0, \ldots, n$.

k	Possible values of T_+
0	0
1	1,2,...,5
2	1+2=3, 1+3=4, 1+4=5, 2+3=5, 1+5=6, 2+4=6, 2+5=7, 3+4=7, 3+5=8, 4+5=9
3	1+2+3=6, 1+2+4=7, 1+2+5=8, 1+3+4=8, 1+3+5=9, 2+3+4=9, 1+4+5=10, 2+3+5=10, 2+4+5=11, 3+4+5=12
4	1+2+3+4=10, 1+2+3+5=11, 1+2+4+5=12, 1+3+4+5=13, 2+3+4+5=14
5	1+2+3+4+5=15

Example: $n = 5$

Therefore the probability distribution of T_+ given k is given by

	Possible values of T_+															
k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0	0	0	0	0	0	0	0	0	0
2	0	0	0	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	0	0	0	0	0	0
3	0	0	0	0	0	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Example: $n = 5$

Hence, using the previous formula, we can compute $P[T_+ = t]$:

	Possible values of T_+							
t	0	1	2	3	4	5	6	7
$P[T_+ = t]$	0.031	0.031	0.031	0.062	0.062	0.094	0.094	0.094
$P[T_+ \leq t]$	0.031	0.062	0.094	0.156	0.219	0.312	0.406	0.500
t	8	9	10	11	12	13	14	15
$P[T_+ = t]$	0.094	0.094	0.094	0.062	0.062	0.031	0.031	0.031
$P[T_+ \leq t]$	0.594	0.688	0.781	0.844	0.906	0.938	0.969	1.000

Note: The same calculation WITH zeros

If the $x_i = 0$ data are left in the sample, then the calculation becomes more complicated

- The selection of the test statistic is not straightforward; T_+ and T_- are still the obvious choices that will distinguish H_a from H_0.
- In the presence of zeros the distribution of K is **no longer Binomial**.
- We need to propose a model for the **number** of zeros.
- We would need a **different** table of critical values for each different number of zeros.

Overall, it is more straightforward to **omit** the zeros; it can be shown that this does not compromise the effectiveness of the test.