STA 104 Applied Nonparametric Statistics

Chapter 2: One-Sample Methods for Location Problem

Xiner Zhou

Department of Statistics, University of California, Davis

Table of contents

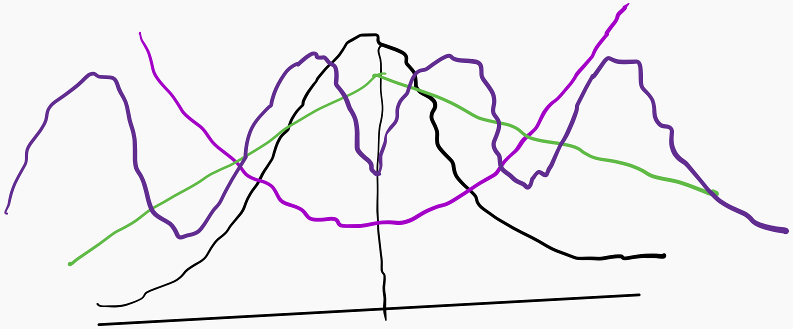
Wilcoxon Signed Rank Test

Setting

Suppose we have a random sample $x_1 \ldots x_n$, i.e. data

Assumptions

- The x's are mutually independent.
- they are from a population that is
  - continuous
  - symmetric about the median θ



Hypothesis

Two-Sided Test:

$$H_0 : \theta = \theta_0 \text{ versus } H_a : \theta \neq \theta_0$$

One-Sided Upper-Tail Test:

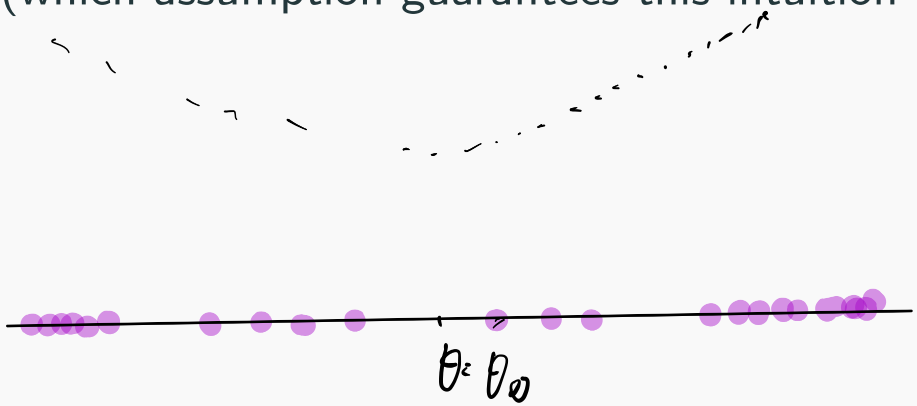
$$H_0 : \theta = \theta_0 \text{ versus } H_a : \theta > \theta_0$$

One-Sided Lower-Tail Test:

$$H_0 : \theta = \theta_0 \text{ versus } H_a : \theta < \theta_0$$

Motivation

Intuition: If θ_0 was the true median of the population, then the magnitude in terms of absolute value of the centered data is nothing to do with the sign of the centered data (which assumption guarantees this intuition valid?)



Assume : Symmetric dist.

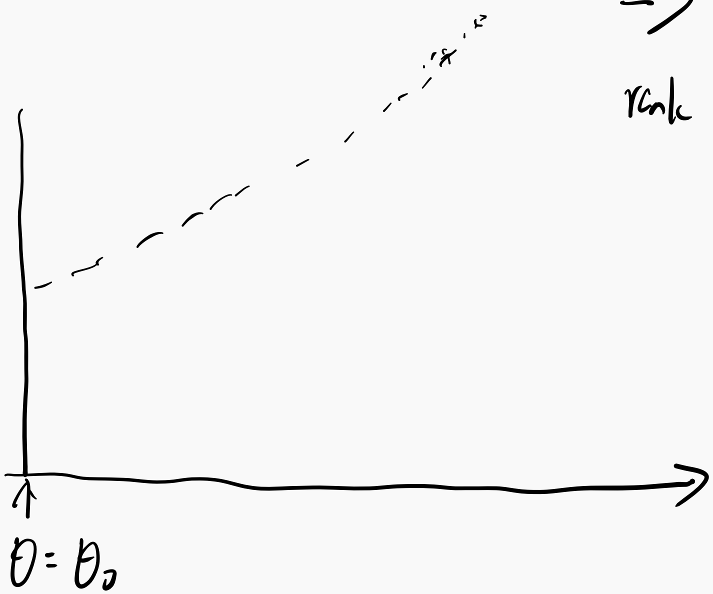
Motivation

- **Centering:** subtract θ_0 from each observation $x_1, \ldots, x_n$ to form a modified sample $\boxed{x_1' = x_1 - \theta_0, \ldots, x_n' = x_n - \theta_0}$ centered data
- **Flip:** form absolute values $|x_1'| \ldots |x_n'|$
- **Rank:** Order them from least to greatest, let R_i denote the rank of ith observation $|x_i'|$

1 3 4 2

$\Rightarrow$ 1 2 3 4

rank : 1 2 3 4



Motivation

- **Sign:**

X_i'

Define indicator variable for signs $\psi_i = \begin{cases} 1, & \text{if } Z_i > 0 \\ 0, & \text{if } Z_i < 0 \end{cases}$

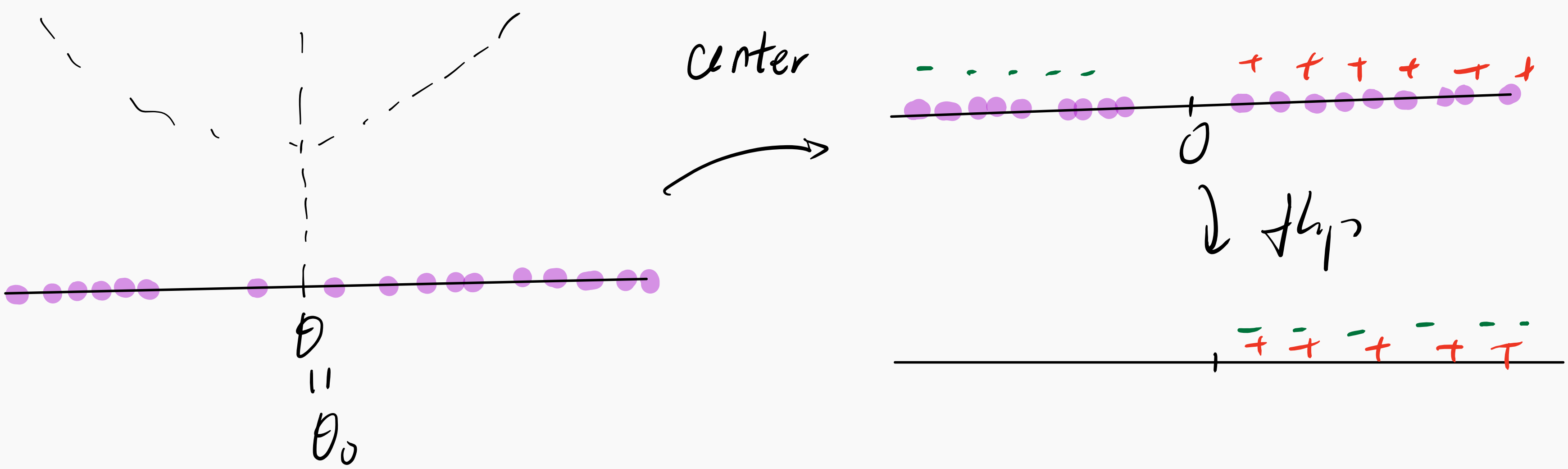
X_i' X_i'

Define positive signed rank of ith observation $\psi_i R_i = \begin{cases} R_i, & \text{if } Z_i > 0 \\ 0, & \text{if } Z_i < 0 \end{cases}$

X_i'

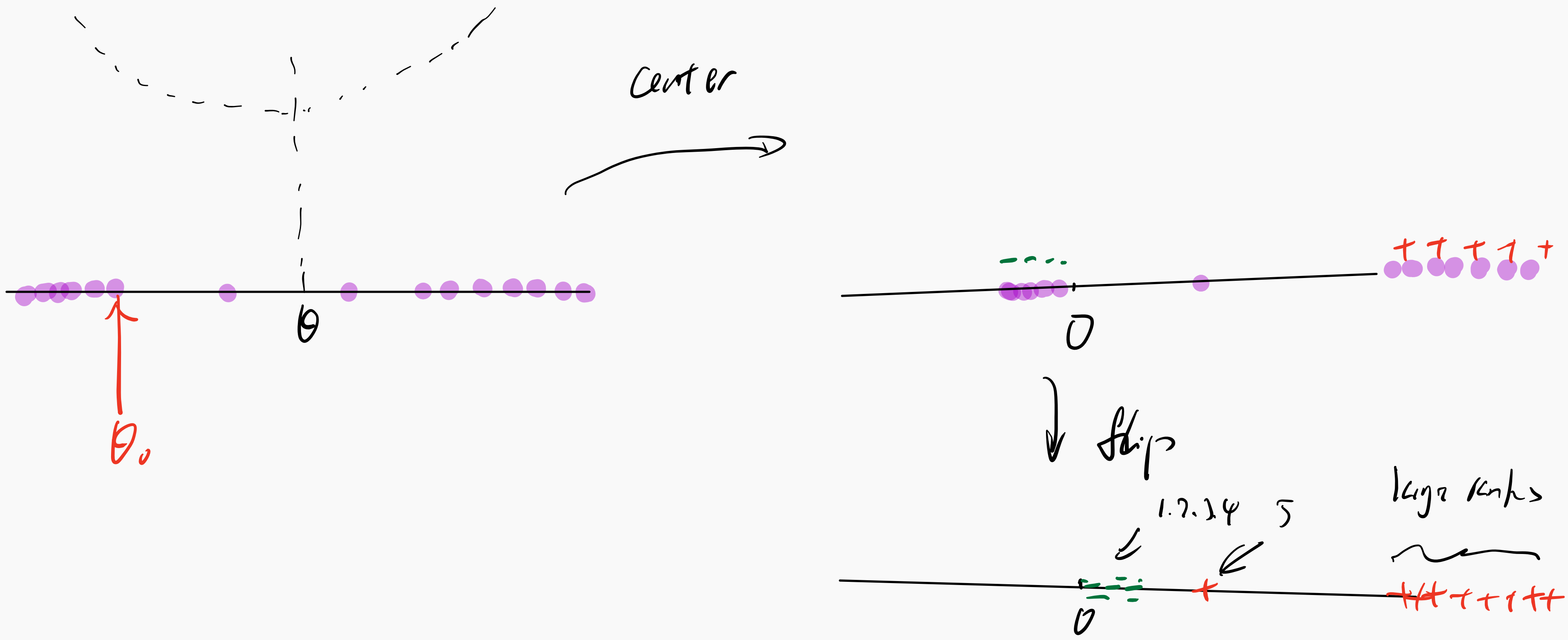
Motivation

- When the true unknown median θ is equal to the hypothesized value θ_0:
  - the centered data will tend to be spread symmetrically around 0, and roughly half of the observations have positive signs, due to symmetric underlying distribution,
  - the ranks associated with negative and positive observations are roughly equal,
  - so the total ranks of those positive signed observation roughly is half of the total ranks of all observations.



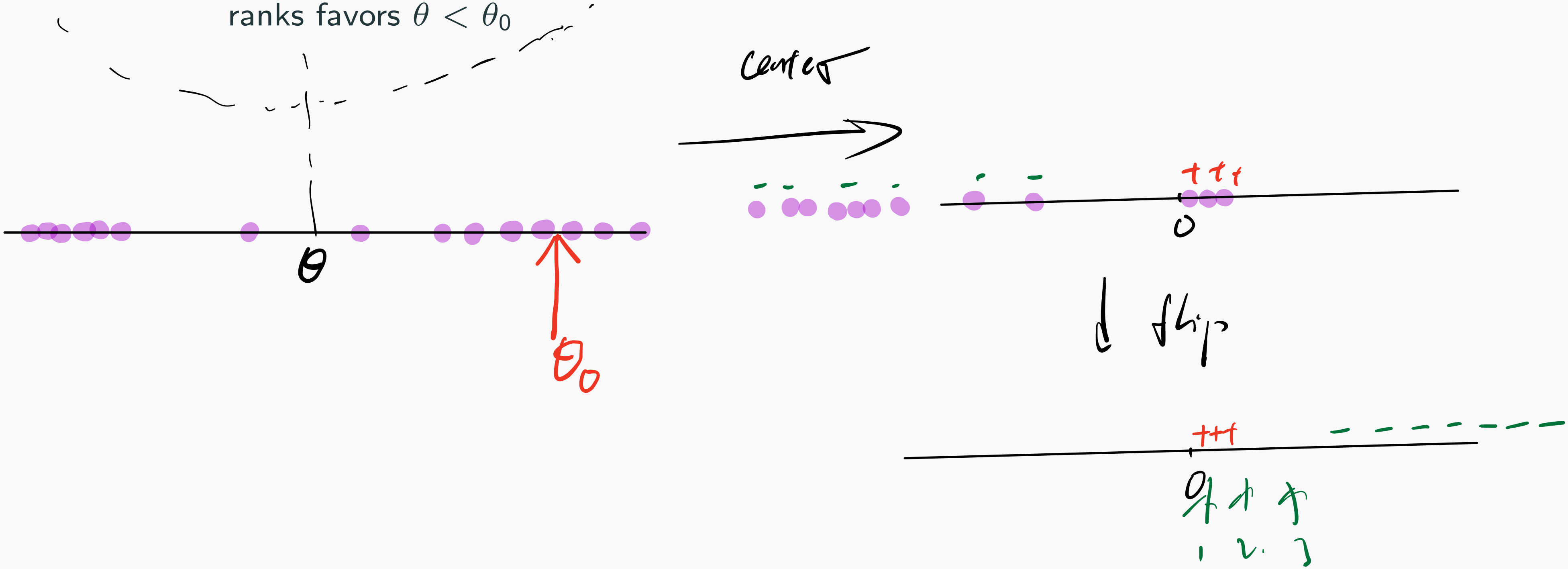
Motivation

- When the true unknown median $\theta > \theta_0$:
  - there will tend to be a larger portion of observations have positive signs and larger ranks associated with them,
  - so the total ranks of those positive signed observation is larger, so large total signed ranks favors $\theta > \theta_0$



Motivation

- When the true unknown median $\theta < \theta_0$:
  - there will tend to be a smaller portion of observations have positive signs and smaller ranks associated with them,
  - so the total ranks of those positive signed observation is smaller, so small total signed ranks favors $\theta < \theta_0$



Motivation

Define Wilcoxon signed rank statistics

$$T^+ = \underbrace{\sum_{i=1}^{n} \psi_i R_i}_{\text{sum of positive signed ranks}}$$

Derivation of (exact) null distribution

When $H_0: \theta = \theta_0$ is true:

- Think the process as randomly split ranks $1, 2, 3, \dots, n$ into two groups
- T^+ is sum of one of groups

Because ranks and signs are independent under H_0

$\Rightarrow$ each rank is equally likely to be + or − : $P(\psi_i = 1) = \frac{1}{2}$

$\Rightarrow$ each configuration of positive signed ranks $\{1\psi_1, 2\psi_2 \dots n\psi_n\}$ occurs with $\left(\frac{1}{2}\right)^n$

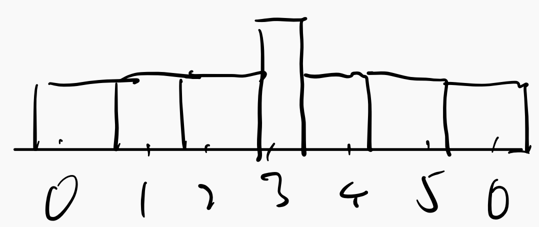
Example

$n=3$

$(1\Psi_1, 2\Psi_2, 3\Psi_3)$	Prob. under H_0	T^+
$(0, 0, 0)$	$(\frac{1}{2})^3 = \frac{1}{8}$	0
$(1, 0, 0)$	$\frac{1}{8}$	1
$(0, 2, 0)$	$\frac{1}{8}$	2
$(0, 0, 3)$	$\frac{1}{8}$	3
$(1, 2, 0)$	$\frac{1}{8}$	3
$(1, 0, 3)$	$\frac{1}{8}$	4
$(0, 2, 3)$	$\frac{1}{8}$	5
$(1, 2, 3)$	$\frac{1}{8}$	6

$\Rightarrow$ null dist. of T^+

T^+	prob
0	1/8
1	1/8
2	1/8
3	2/8
4	1/8
5	1/8
6	1/8



We have derived the null distribution of T^+ without specifying the forms of the underlying populations beyond the point of requiring that they be continuous and symmetric about zero. This is why the test procedures based on T^+ are called distribution-free procedures.

From the null distribution of T^+we can determine the critical value t_α and control the probability α of falsely rejecting H_0 when H_0 is true.

■ Wilcoxon signed rank test statistic

data: -2 -1.5 -1 0.9 1.1 1.6 2.2

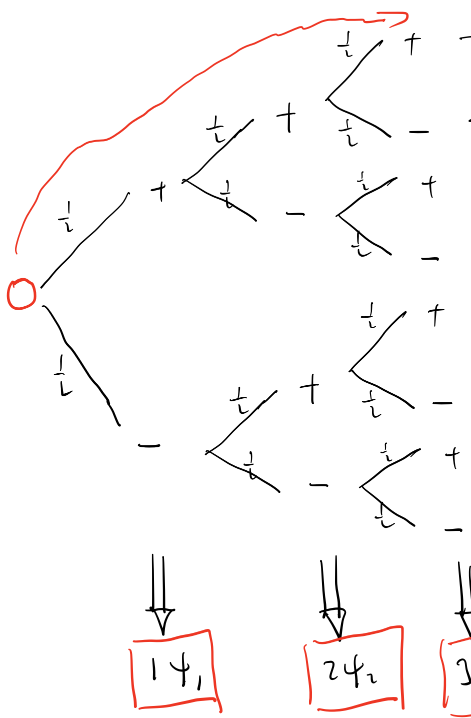
hypothesized median 0

rank: 6 4 2 1 3 5 7

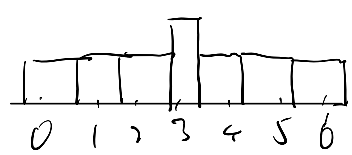
$\Rightarrow T^+ = 1+3+5+7 = 16$

■ (exact) null distribution of T^+

key: each rank 1. 2. 3 ... n has equal probability to have + or - sign



$(1\psi_1, 2\psi_2, 3\psi_3)$	T^+	Prob
(1, 2, 3)	6	$\frac{1}{8}$
(1, 2, 0)	3	$\frac{1}{8}$
(1, 0, 3)	4	$\frac{1}{8}$
(1, 0, 0)	1	$\frac{1}{8}$
(0, 2, 3)	5	$\frac{1}{8}$
(0, 2, 0)	2	$\frac{1}{8}$
(0, 0, 3)	3	$\frac{1}{8}$
(0, 0, 0)	0	$\frac{1}{8}$



null dist. of T^+

distribution-free property

= nonparametric

Large sample approximation of null distribution:

Reorder the data according to their absolute ranks, let $V_i = i\cancel{S_i}$ be the ith observation.

$i\psi_i$

$$T^+ = \sum_{i=1}^{n} \psi_i R_i$$

$$= \sum_{i=1}^{n} V_i$$

V_i are mutually independent dichotomous random variables with $P(V_i = i) = P(V_i = 0) = \frac{1}{2}$. T^+ is sum of independent random variables, follows from standard theory for sums of mutually independent, but not identically distributed, random variables, such as the Liapounov central limit theorem (cf. Randles and Wolfe (1979, p. 423)), it has an asymptotic normality distribution.

Define standardized Wilcoxon signed rank statistics

$$\Rightarrow \text{ standardized form } T^* = \frac{T^+ - E_0\left(T^+\right)}{\left\{\mathrm{var}_0\left(T^+\right)\right\}^{1/2}} = \frac{T^+ - \frac{n(n+1)}{4}}{\left\{\frac{n(n+1)(2n+1)}{24}\right\}^{1/2}}$$

$$\sim N(0,1)$$

if n is large

n: sample size

Optimal :

(1)

$$E_0\left(T^+\right) = E\left[\sum_{i=1}^{n} V_i\right] = \sum_{i=1}^{n} E\left[V_i\right]$$

$$E_0\left(V_i\right) = i\left(\frac{1}{2}\right) + 0\left(\frac{1}{2}\right) = \frac{i}{2}$$

$$\Rightarrow E_0\left(T^+\right) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{1}{2}\left[\frac{n(n+1)}{2}\right] = \boxed{\frac{n(n+1)}{4}}$$

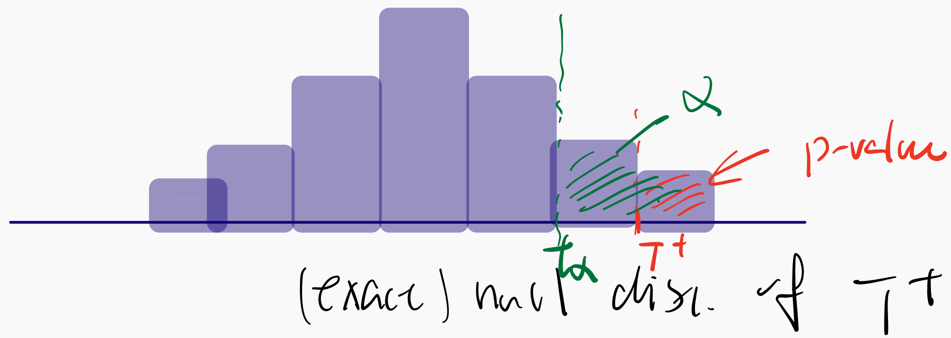
(2)

$$\text{var}_0\left(T^+\right) = \text{var}\left(\sum_{i=1}^{n} V_i\right) = \sum_{i=1}^{n} \text{var}\left(V_i\right)$$

$$\begin{aligned}
\text{var}_0\left(V_i\right) &= E_0\left(V_i^2\right) - \left[E_0\left(V_i\right)\right]^2 \\
&= \left[i^2\left(\frac{1}{2}\right) + 0^2\left(\frac{1}{2}\right)\right] - \left[\frac{i}{2}\right]^2 \\
&= \frac{i^2}{2} - \frac{i^2}{4} = \frac{i^2}{4}
\end{aligned}$$

$$\Rightarrow \text{var}_0\left(T^+\right) = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{1}{4}\left[\frac{n(n+1)(2n+1)}{6}\right] = \boxed{\frac{n(n+1)(2n+1)}{24}}$$

Procedure



a. One-Sided Upper-Tail Test.

To test

$$H_0 : \theta = \cancel{0}\ \theta_0$$

versus

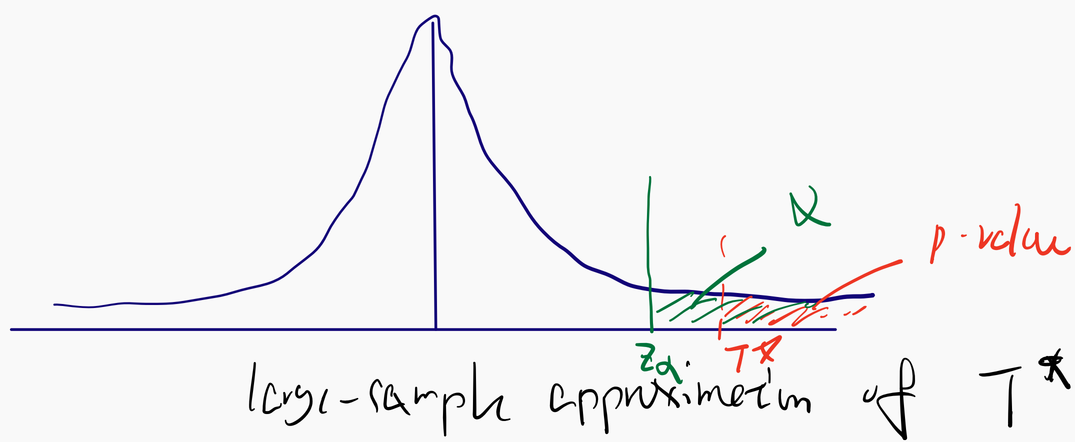
$$H_a : \theta > \cancel{0}\ \theta_0$$

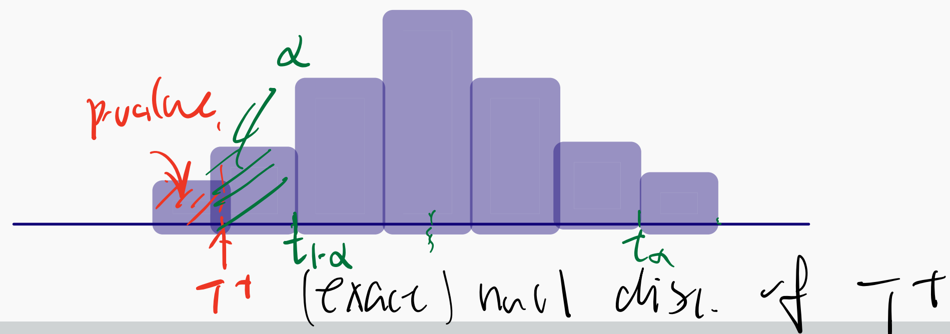
Large T^+ / T^* favor H_a

at the α level of significance, Reject H_0 if $T^+ \geq t_\alpha$; otherwise do not reject.

The normal approximation:

Reject H_0 if $T^* \geq z_\alpha$; otherwise do not reject.





b. One-Sided Lower-Tail Test.

To test

$$H_0 : \theta = \cancel{0}\ \theta_0$$

versus

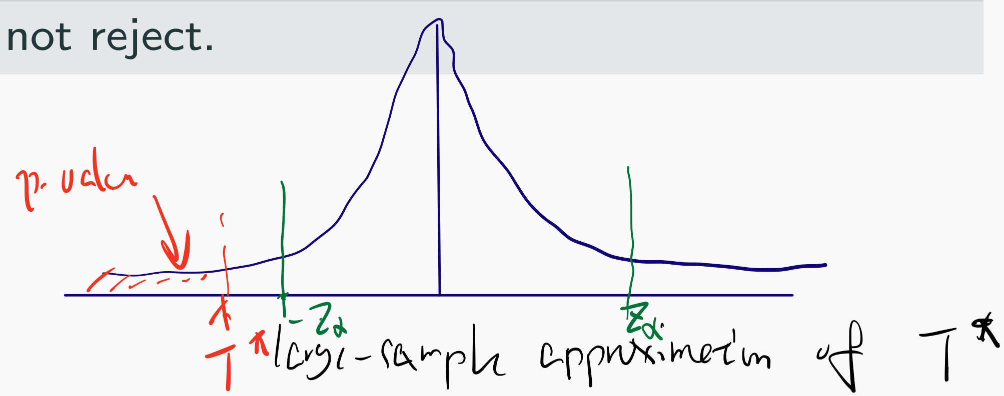
$$H_a : \theta < \cancel{0}\ \theta_0$$

Small T^+ / T^* favor H_a

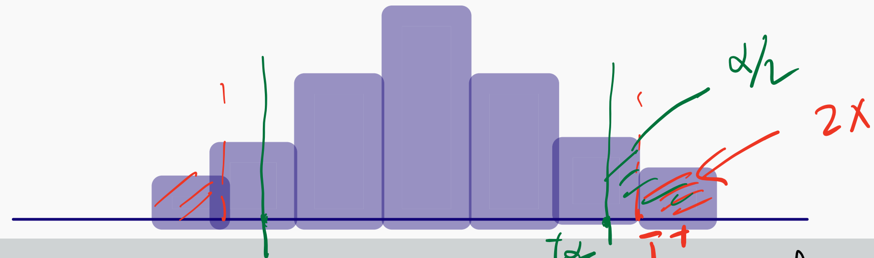
at the α level of significance, Reject H_0 if $T^+ \le \boxed{t_{1-\alpha}} = \frac{n(n+1)}{2} - \boxed{t_\alpha}$; otherwise do not reject.

The normal approximation:

Reject H_0 if $T^* \le \underbrace{-z_\alpha}_{= z_{1-\alpha}}$; otherwise do not reject.



Procedure



$t_{1-\frac{\alpha}{2}}$ (exact) null dist. of T^+

c. Two-Sided Test.

To test

$$H_0 : \theta = \cancel{0}\ \theta_0$$

versus

$$H_a : \theta \neq \cancel{0}\ \theta_0$$

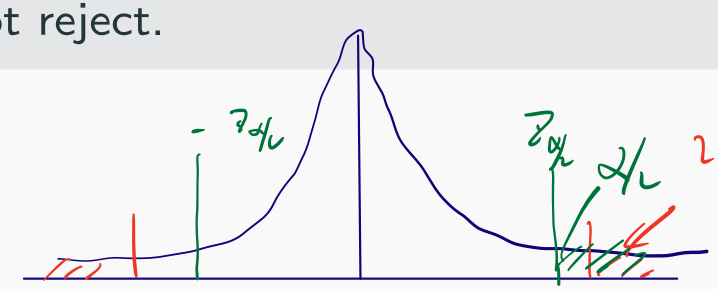
Both small and large T^+ / T^* favor H_a

at the α level of significance, Reject H_0 if $T^+ \geq t_{\alpha/2}$ or $T^+ \leq t_{1-\alpha/2} = \frac{n(n+1)}{2} - t_{\alpha/2}$; otherwise do not reject.

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of T^+.

The normal approximation:

Reject H_0 if $|T^*| \geq z_{\alpha/2}$; otherwise do not reject.



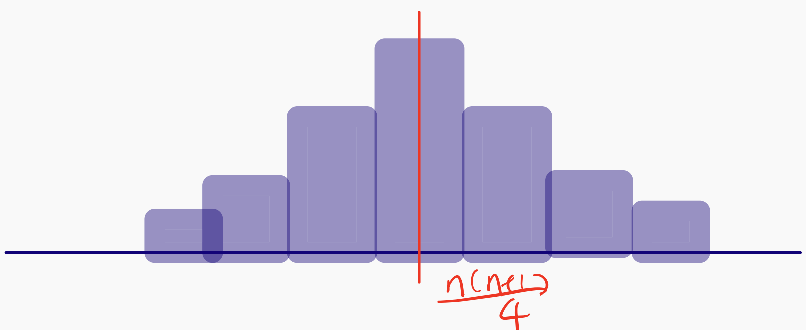
large-sample approximation of T^*

An estimator associated with Wilcoxon's signed rank statistics (Hodges-Lehmann)

The null distribution of the statistic T^+ is symmetric about its mean, $n(n+1)/4$. A natural estimator of θ is the amount that should be subtracted from each Z_i so that the value of T^+, when applied to the shifted sample $X_1 - \hat{\theta}, \ldots, X_n - \hat{\theta}$, is as close to $n(n+1)/4$ as possible.

Roughly speaking, we estimate θ by the amount ($\hat{\theta}$) that the Z sample should be shifted in order that $Z_1 - \hat{\theta}, \ldots, Z_n - \hat{\theta}$ appears (when "viewed" by the signed rank statistic T^+) as a sample from a population with median 0.

If we have guessed correct median :



(exact) null dist. of T^+

Optional:

$$T^{+} = \sum_{i=1}^{n} S_i R_i$$

$$\Rightarrow \frac{1}{\frac{n(n+1)}{2}} \sum I\left(\frac{X_i + X_j}{2} \geq \theta_0\right)$$

$$\approx \frac{1}{2}$$

$$\psi_i = I\left(X_i > \theta_0\right)$$

$$R_i = \sum_{j=1}^{n} I\left(\left|X_j - \theta_0\right| \leq \left|X_i - \theta_0\right|\right)$$

	X_1	X_2	X_3
X_1	$\frac{X_1+X_1}{2}$	$\frac{X_1+X_2}{2}$	$\frac{X_1+X_3}{2}$
X_2		$\frac{X_2+X_2}{2}$	$\frac{X_2+X_3}{2}$
X_3			$\frac{X_3+X_3}{2}$

re-arrange

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} I\left(\left|X_j - \theta_0\right| \leq \left|X_i - \theta_0\right|, X_i > \theta_0\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} I\left(\left|X_j - \theta_0\right| \leq X_i - \theta_0, X_i > \theta_0\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} I\left(\theta_0 - X_i \leq X_j - \theta_0 \leq X_i - \theta_0\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} I\left(2\theta_0 \leq X_j + X_i \leq 2X_i\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} I\left(\frac{X_i + X_j}{2} \geq \theta_0, X_j \leq X_i\right)$$

$$= \boxed{\sum_{1 \leq i \leq j \leq n} 1\left(\frac{X_i + X_j}{2} \geq \theta_0\right)} \approx \frac{n(n+1)}{4}$$

$\Rightarrow$ The median θ_0 should be the value that splits the pairwise averages into 50% / 50%.

Procedure

The Walsh Averages.

Each of the $n(n+1)/2$ averages $\left(X_i + X_j\right)/2$, $i \le j = 1, \ldots, n$, is called a Walsh average.

Estimate

To estimate the median θ, form the $M = n(n+1)/2$ averages $\left(X_i + X_j\right)/2$, for $i \le j = 1, \ldots, n$. The estimator of θ associated with the Wilcoxon signed rank statistic T^+ is

$$\hat{\theta} = \text{median}\left\{\frac{\cancel{X}_i + \cancel{X}_j}{2}, i \le j = 1, \ldots, n\right\}.$$

Let $W^{(1)} \le \cdots \le W^{(M)}$ denote the ordered values of $\left(X_i + X_j\right)/2$, where $M = \frac{n(n+1)}{2}$

- Then if M is odd, say $M = 2k+1$, we have $k = (M-1)/2$ and

  $$\hat{\theta} = W^{(k+1)},$$

  the value that occupies position $k+1$ in the list of the ordered $\left(X_i + X_j\right)/2$ averages.
- If M is even, say $M = 2k$, then $k = M/2$ and

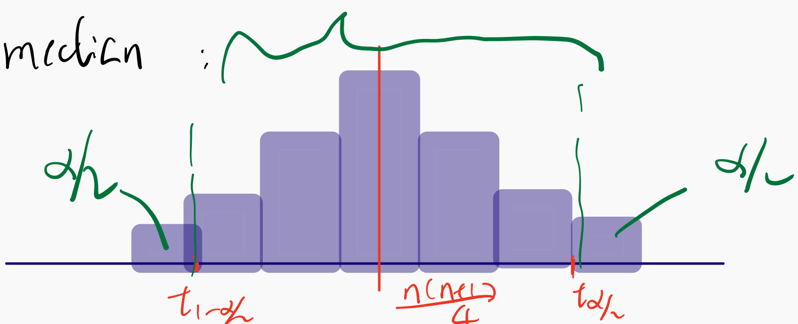
  $$\hat{\theta} = \frac{W^{(k)} + W^{(k+1)}}{2}$$

  That is, when M is even, $\hat{\theta}$ is the average of the two $\left(X_i + X_j\right)/2$ values that occupy positions k and $k+1$ in the ordered list of the M $\left(X_i + X_j\right)/2$ averages. The $\left(X_i + X_j\right)/2$.

The true population median θ_0 is the value such that the number of Walsh averages above it is the Wilcoxon's signed rank statisitcs T^+ which should be centered at $\frac{n(n+1)}{4}$ with some natural variation. So we use the natural variation of T^+ reverse engineer the most probable region of θ_0 .

$$T^+ = \underbrace{\sum_{1 \le i \le j \le n} I\left(\frac{X_i + X_j}{2} \ge \theta_0\right)}_{\text{total number of Walsh averages above true median } \theta_0}$$

If we have guessed correct median :



(exact) null dist. of T^+

$$\underbrace{P\left(\boxed{t_{1-\alpha/2}} \le T^+ \le \boxed{t_{\alpha/2}}\right) = \boxed{1-\alpha}}_{\text{with probability } (1-\alpha)100\%} \approx 95\%$$

$\Rightarrow$ total number of Walsh averages above true median should be between $t_{1-\alpha/2}$ and $t_{\alpha/2}$

Procedure

$W^{(1)}$ . . . $W^{\left(\frac{n(n+1)}{2}\right)}$

$\uparrow$ $W^{\left(\frac{n(n+1)}{2}+1-t_{\alpha/2}\right)}$ $\uparrow$ $W^{(t_{\alpha/2})}$

$(1-\alpha)100\%$ Confidence Interval

For a symmetric two-sided confidence interval for median θ, with confidence coefficient $1-\alpha$, set

$$\cancel{t_\alpha = \frac{n(n+1)}{2} + 1 - t_{\alpha/2},}$$

where $t_{\alpha/2}$ is the upper $(\alpha/2)$ th percentile point of the null distribution of T^+.

The $100(1-\alpha)\%$ confidence interval (θ_L, θ_U) for θ that is associated with the Wilcoxon signed rank statistics is

$$W\left(\frac{n(n+1)}{2}+1-t_{\alpha/2}\right)$$

$$\theta_L = \cancel{W^{(t_\alpha)}}, \quad \theta_U = \cancel{W^{(M+1-C_\alpha)}} = W^{(t_{\alpha/2})}$$

~~where $M = n(n+1)/2$ and $W^{(1)} \leq \cdots \leq W^{(M)}$ are the ordered values of the $(X_i + X_j)/2$ Walsh averages. θ_L is the Walsh average that occupies position t_α in the list. The upper end point θ_U is the Walsh average that occupies the position $M + 1 - t_\alpha = t_{\alpha/2}$ in this ordered list.~~

$W^{(1)}$ $W^{(2)}$ $W^{(3)}$ $W^{(4)}$ $W^{(5)}$ $W^{(6)}$ $W^{(7)}$ $W^{(8)}$

$t_{\alpha/2} = 7$

$8+1-7=2$

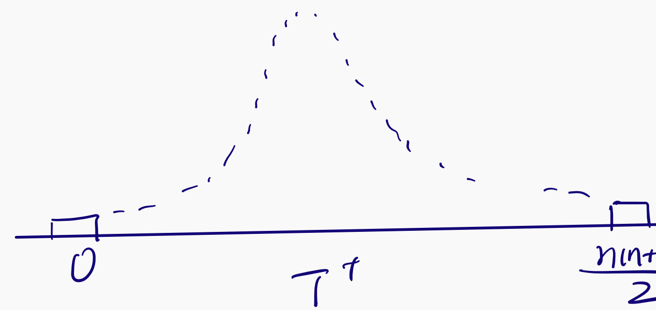
$\Rightarrow$ C.I.: $(\theta_L = W^{(2)}, \ \theta_U = W^{(7)})$

Example: The Mariner and the Pioneer Spacecraft Data

The data were reported by Anderson, Efron, and Wong (1970). The seven observations represent average measurements of, the ratio of the mass of the Earth to that of the moon, obtained from seven different spacecraft.

On the basis of the previous (2-3 years earlier) Ranger spacecraft findings, scientists had considered the value of the ratio of the mass of the Earth to that of the moon to be approximately 81.3035. Thus, we are interested in testing $H_0 : \theta = 81.3035$ versus the alternative $\theta \neq 81.3035$.

		Step 1	Step 2	Step 3
i	X_i	$X_i' = X_i - 81.3035$	Sign (ψ_i)	Rank (R_i)
1	81.3001	$-.0034$	$-$	6
2	81.3015	$-.0020$	$-$	2
3	81.3006	$-.0029$	$-$	4
4	81.3011	$-.0024$	$-$	3
5	81.2997	$-.0038$	$-$	7
6	81.3005	$-.0030$	$-$	5
7	81.3021	$-.0014$	$-$	1



$$\frac{n(n+1)}{2} = \frac{7 \times 8}{2} = 28$$

Exact test:

$$T^+ = 0$$

$$p - value = 1/2^7 = 0.015625$$

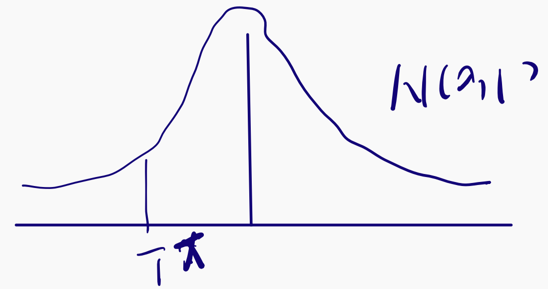
$$P(T^+ \leq 0 \text{ or } T^+ \geq 28) = 2P(T^+ \leq 0)$$
$$= 2P(T^+ = 0)$$
$$= 2 \times \left(\frac{1}{2}\right)^7$$

Confirm with built-in function:

```
> wilcox.test(c(81.3001,81.3015,81.3006,81.3011,81.2997,81.3005,81.3021),mu=81.3035)

        Wilcoxon signed rank exact test

data:  c(81.3001, 81.3015, 81.3006, 81.3011, 81.2997, 81.3005, 81.3021)
V = 0, p-value = 0.01563
alternative hypothesis: true location is not equal to 81.3035
```



Large-sample approximation:

$$T^* = \frac{T^+ - \frac{n(n+1)}{4}}{\left\{\frac{n(n+1)(2n+1)}{24}\right\}^{1/2}} = \frac{0 - [7(8)/4]}{[7(8)(15)/24]^{1/2}} = -2.366$$

$$p - value = 0.01798144 \Leftarrow 2P(\underbrace{Z}_{\sim N(0,1)} \leq T^*)$$

> pnorm( ) in R

Both the exact test and the large-sample approximation indicate the existence of strong evidence to reject the findings of the earlier Ranger spacecraft that $\theta = 81.3035$.

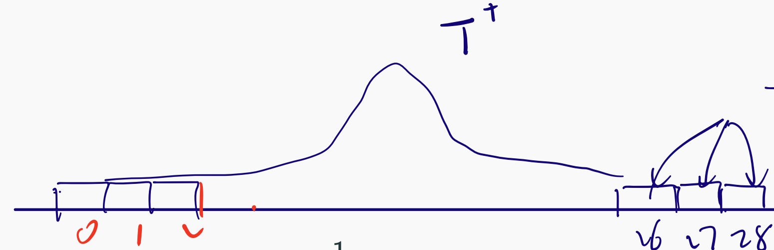
Welsh averages	x_1	x_2	x_3	x_4
x_1				
x_2				
x_3				
x_4				

An estimate for median:

```
> library(Rfit)
> sort(walsh(c(81.3001,81.3015,81.3006,81.3011,81.2997,81.3005,81.3021)))
 [1] 81.29970 81.29990 81.30010 81.30010 81.30015 81.30030 81.30035 81.30040
 [9] 81.30050 81.30055 81.30060 81.30060 81.30060 81.30080 81.30080 81.30085
[17] 81.30090 81.30100 81.30105 81.30110 81.30110 81.30130 81.30130 81.30135
[25] 81.30150 81.30160 81.30180 81.30210
```

$$M = 7(8)/2 = 28, \text{ we see that } M = 2k \text{ with } k = 14$$

$$\Rightarrow \hat{\theta} = \frac{W^{(14)} + W^{(15)}}{2} = \frac{81.3008 + 81.3008}{2} = 81.3008$$



$\frac{1}{2^7} = 0.0078125$

95% Confidence interval for median:

With $n = 7$ and $\alpha = .05$, each configuration under null has equal probability of $\frac{1}{2^7} = 0.0078125$, there should be at most 3.2 configurations to the right of $t_{\alpha/2} = 26$. Thus, $\cancel{t_{1-\alpha/2}} = 28 + 1 - 26 = 3$.

$= \frac{0.025}{0.0078125}$

$t_{\alpha/2} = t_{0.025}$

$$\theta_{\mathrm{L}} = W^{(3)} = 81.3001 \text{ and } \theta_{\mathrm{U}} = W^{(26)} = 81.3016$$

so that our 95% confidence interval for θ is

$$(\theta_{\mathrm{L}}, \theta_{\mathrm{U}}) = (81.3001, 81.3016)$$

$W^{(1)}$ $W^{(2)}$ $W^{(3)}$

$W^{(26)}$ $W^{(27)}$ $W^{(28)}$

Confirm with built-in function:

Correction

```
> wilcox.test(c(81.3001,81.3015,81.3006,81.3011,81.2997,81.3005,81.3021),
m=28, m=81.3035, exact=T, conf.int=T,conf.level=0.95)
```

mu

```
        Wilcoxon signed rank exact test

data:  c(81.3001, 81.3015, 81.3006, 81.3011, 81.2997, 81.3005, 81.3021)
V = 28, p-value = 0.01563
alternative hypothesis: true location is not equal to 28
95 percent confidence interval:
 81.3001 81.3016
sample estimates:
(pseudo)median
       81.3008
```