Nonparametric Statistics

Ch.2 One-sample (location) problem

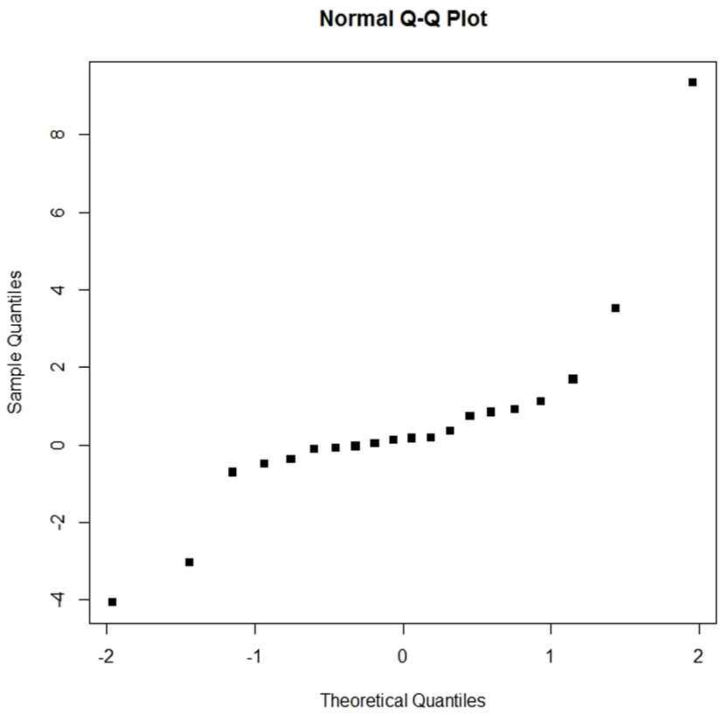
Motivating example

Ex1] A random sample with size 20 was observed:

(1.12 0.84 -0.71 -0.38 -4.05 -0.08 1.68 -0.49 0.17 3.51 -0.04 0.19 0.13 0.91 -0.11 0.74 9.33 0.35 -3.04 0.04).

Then, $\bar{X} = 0.51,\ S = 2.59$

Is it possible to conduct the t-test to see if the population mean is zero? In other words, does the random sample come from a normal distribution?



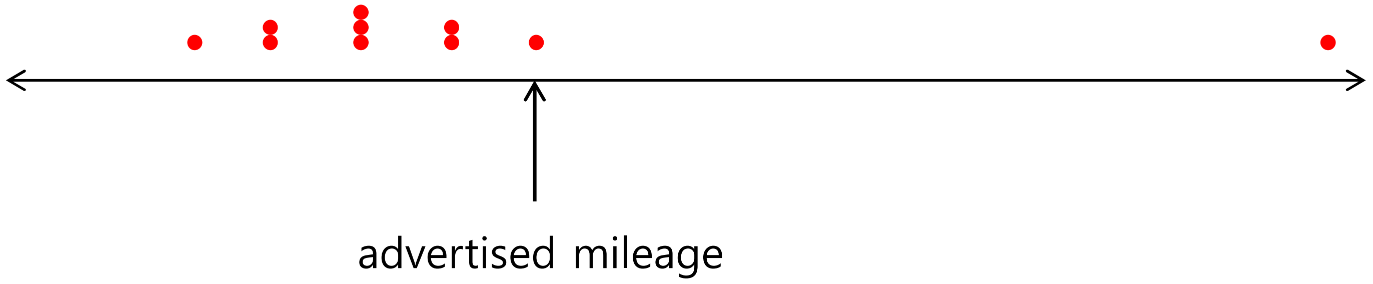
→ Clear evidence for the violation of the normality assumption

Motivating example

Ex2] A consumer group wishes to see if the actual mileage of a new car matches the advertised 17 miles per gallon. The group records the mileage by ten times repetition.

(14, 15, 13, 17, 16, 16, 15, 25, 14, 15)

In this example, the sample mean is 16, and p-value for the one-sided t-test is 0.1861. Do you think the advertisement of the company is true?



Review : One-sample t-test

Let $X_1, \dots, X_n$ be a random sample from $N(\mu, \sigma^2)$. A test for the hypothesis

$$H_0: \mu = \mu_0 \qquad vs \qquad H_1: \mu \neq \mu_0$$

can be performed with the test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

- T has the t-distribution with n-1 degrees of freedom under H_0.
- Rejection region for the test at the significance level α is $\{t: |t| > t_{\alpha/2}\}$ where $t_{\alpha/2}$ satisfies $P(T > t_{\alpha/2})$= $\alpha/2$ with $T \sim T(n-1)$.
- p-value is computed by $P(|T| > |t_0|)$ where t_0 is the observed value of the test statistic.

Note] In the case that the alternative is one-sided like $H_1: \mu > \mu_0$ or $H_1: \mu < \mu_0$, the rejection region and the p-value will be given as:

$$H_1: \mu > \mu_0 \quad : \quad \{t: t > t_\alpha\} \text{ and } P(T > t_0)$$

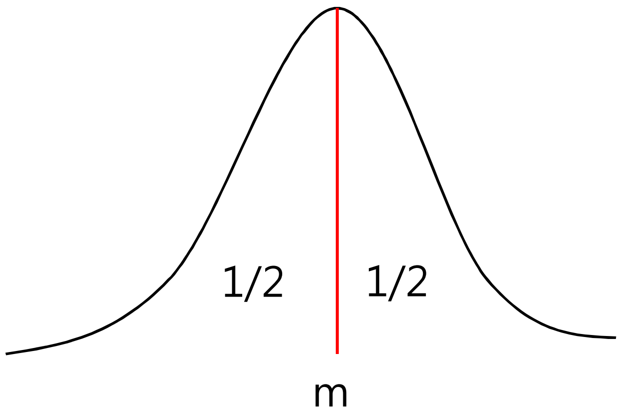
$$H_1: \mu < \mu_0 \quad : \quad \{t: t < t_\alpha\} \text{ and } P(T < t_0)$$

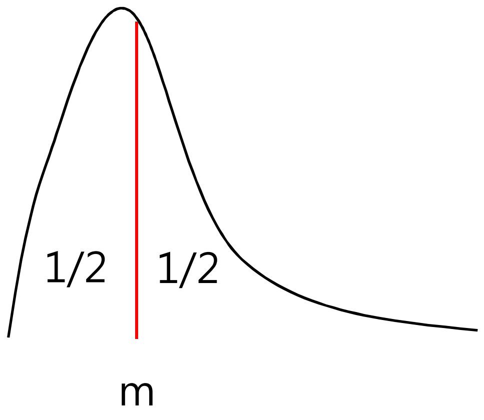
Review : One-sample t-test

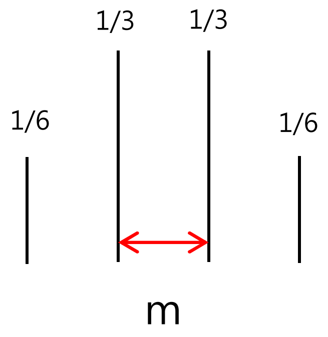
- ❖ One-sample t-test is the test for the mean.
- ❖ The mean represents the central location of a distribution.
- ❖ As for the normal distribution, the mean coincides with the median.
- ❖ Note] The median (m) of a distribution (or a random variable X) is the value separating the higher half of the distribution from the lower half. It satisfies

$$P(X \leq m) \geq 1/2 \text{ and } P(X \geq m) \geq 1/2.$$

- ❖ If the distribution is continuous, the median is uniquely defined. But, if it is discrete, the median may not be unique.
- ❖ We will assume that the underlying distribution is continuous.







Sign test

When we are suspicious of the normality assumption, one of the alternatives is to use the simple sign test.

Let $X_1, \dots, X_n$ be a random sample from a **continuous** distribution. A test for the hypothesis

$$H_0: m = m_0 \qquad vs \qquad H_1: m \neq m_0$$

where m is the median of the distribution, and therefore satisfies $P(X > m) = P(X < m) = 1/2$,

can be performed with the test statistic

$$S_n = \sum_{i=1}^{n} I(X_i > m_0) \quad , \qquad (I(\cdot) : indicator \quad function \quad).$$

Note] When performing the sign test, delete data values that equal to m_0.

Sign test

- $S_n \sim Binom\ (n, 1/2)$ under H_0.

  $\because I(X_i > \mathrm{m}_0) \sim \mathrm{Bernoulli}(1/2)$ & S_n is the iid sum of them.
- Close values of S_n to 1/2 support H_0. The farther away from 1/2, the stronger evidence we have against H_0.
- Rejection region for the test at the significance level α is

$$\{s: s \geq b(n, 1/2, \alpha/2)\ or\ s \leq b(n, 1/2, \alpha/2) - 1\}$$

where $b(n, 1/2, \alpha/2)$ satisfies $P(B \geq b(n, 1/2, \alpha/2))= \alpha/2$ with $B \sim Binom\ (n, 1/2)$.

- p-value is computed by

$$2 \times P(B \geq \max(s_n, n - s_n))$$

where s_n is the observed value of the test statistic.

Note] In the case that the alternative is one-sided like $H_1: m > m_0$ or $H_1: m < m_0$, the rejection region and the p-value will be given as:

$$H_1: m > m_0 \quad : \quad \{s: s \geq b(n, 1/2, \alpha)\} \text{ and } P(B \geq s_n)$$

$$H_1: m < m_0 \quad : \quad \{s: s \leq b(n, 1/2, 1 - \alpha) - 1\} \text{ and } P(B \leq s_n)$$

Sign test : Example (revisited)

Ex1] A random sample with size 20 was observed:

(1.12 0.84 -0.71 -0.38 -4.05 -0.08 1.68 -0.49 0.17 3.51 -0.04 0.19 0.13 0.91 -0.11 0.74 9.33 0.35 -3.04 0.04).

We want to test

$$H_0 : m = 0 \qquad vs \qquad H_1 : m \neq 0$$

Answer] The number of observations greater than 0 is 12. ∴, the p-value is

$$2 \cdot P(B \geq 12) = 0.5034$$

where $B \sim Biom\ (\mathbf{20}, 1/2)$. We cannot reject the null at the significance level 0.05.

Ex2] A consumer group records the mileage by ten times repetition.

(14, 15, 13, 17, 16, 16, 15, 25, 14, 15)

We want to test

$$H_0 : m = 17 \qquad vs \qquad H_1 : m < 17$$

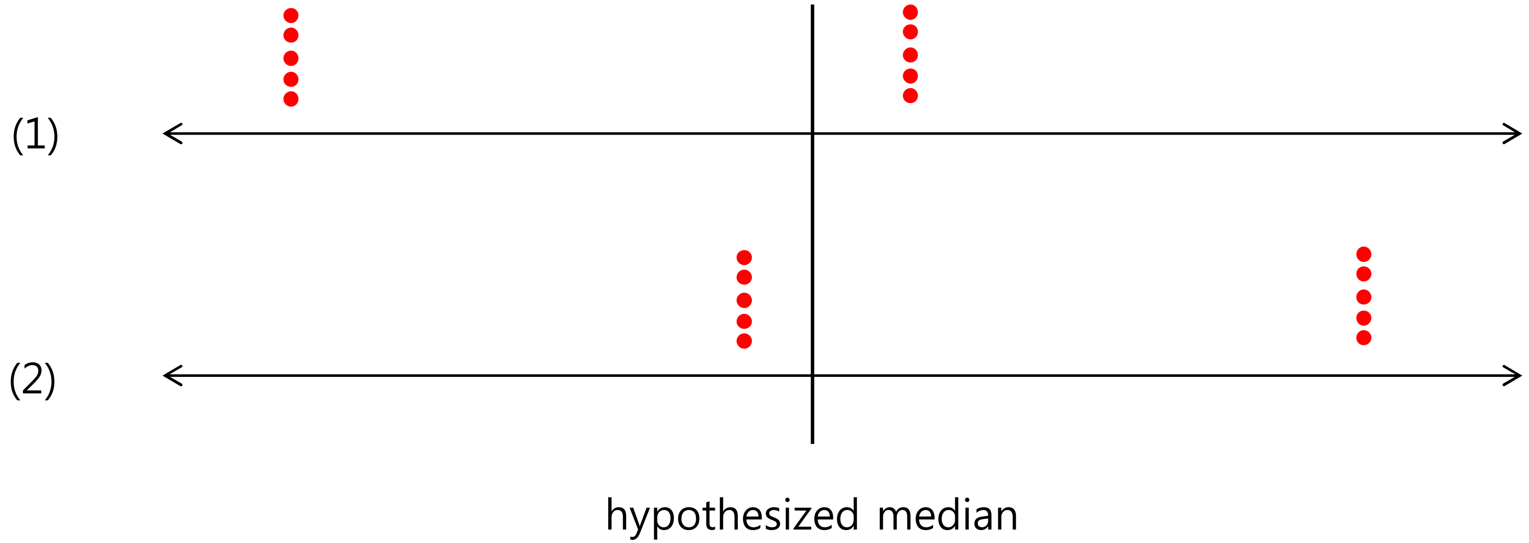
Answer] The number of observations greater than 17 is 1. ∴, the p-value is

$$P(B \leq 1) = 0.0195$$

where $B \sim Biom\ (\mathbf{9}, 1/2)$. We reject the null at the significance level 0.05. Note that the observation 17 was deleted to remove a tie.

Wilcoxon Signed Rank test

The sign test only utilizes the signs. It discards the magnitudes of the differences between the observations and the hypothesized median. Consider the next two sets of data.



These two cases give the same test statistic, 5. But, we can easily see that it is very likely that they actually have different medians.

Wilcoxon Signed Rank test

Let $X_1, \dots, X_n$ be a random sample from a **symmetric continuous** distribution. A test for the hypothesis

$$H_0: m = m_0 \qquad vs \qquad H_1: m \neq m_0$$

where m is the median of the distribution can be performed with the test statistic

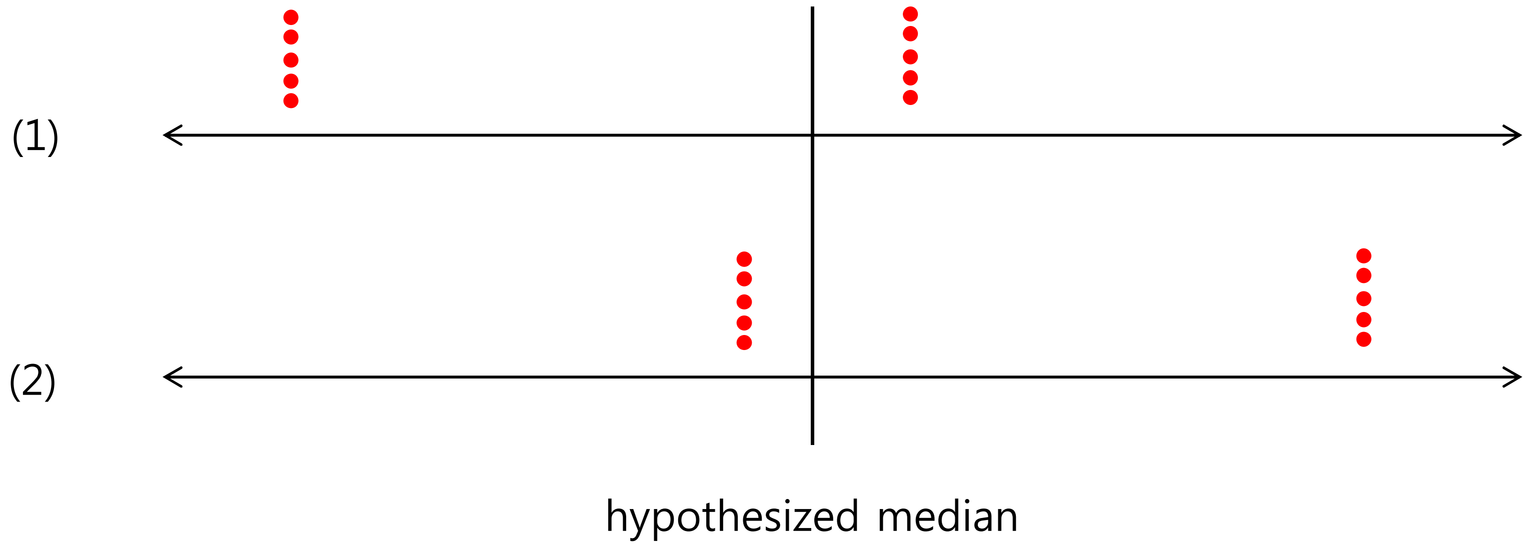
$$W_n = \sum_{i=1}^{n} R_i^+ I(X_i - m_0 > 0) \quad , \qquad (I(\cdot) : indicator \quad function \quad),$$

Where R_i^+ is the rank of the absolute values of $X_i - m_0$.

- If the true median is m_0, it is expected that W_n has a value close to n(n+1)/4 with high probability.
- If the test statistic is far from n(n+1)/4, it implies that the null hypothesis is not true.
- We can compute the exact p-value, but it is too complicated to do it by hand. So, using statistical packages or large sample approximation is recommended.

Wilcoxon Signed Rank test

In the example on page 10,



(1) The observed test statistic is 1+2+3+4+5=15.

(2) The observed test statistic is 6+7+8+9+10=40.

The Wilcoxon signed rank test produces the different results.

Wilcoxon Signed Rank test : Example (revisited)

Ex1] A random sample with size 20 was observed:

(1.12 0.84 -0.71 -0.38 -4.05 -0.08 1.68 -0.49 0.17 3.51 -0.04 0.19 0.13 0.91 -0.11 0.74 9.33 0.35 -3.04 0.04).

We want to test

$$H_0 : m = 0 \qquad vs \qquad H_1 : m \neq 0$$

Answer] The rank of the absolute values are

(15 13 11 9 19 3 16 10 6 18 1.5 7 5 14 4 12 20 8 17 1.5).

The observed test statistic is

$$w_n = 135.5$$

The p-value is given by 0.2627 (exact result computed in **R**).

Wilcoxon Signed Rank test : Example (revisited)

Ex2] A consumer group records the mileage by ten times repetition.

(14, 15, 13, 17, 16, 16, 15, 25, 14, 15)

We want to test

$$H_0: m = 17 \qquad vs \qquad H_1: m < 17$$

Answer]

(1) $X_i - 17$: (-3, -2, -4, -1, -1, -2, 8, -3, -2)

(2) $|X_i - 17|$: (3, 2, 4, 1, 1, 2, 8, 3, 2)

(3) R_i^+ : (6.5, 4, 8, 1.5, 1.5, 4, 9, 6.5, 4)

(4) $w_n = 9$

However, this result is not reliable since the parent distribution does not seem symmetric.

Sign test : large sample approximation

The test statistic for the sign test is

$$S_n = \sum_{i=1}^{n} I(X_i > m_0).$$

Recall that this is the sum of iid random variables so that the **central limit theorem** is applicable.

(1) $E\big(I(X_i > m_0)\big) = P(X_i > m_0) = 1/2$

(2) $\mathrm{Var}\big(I(X_i > m_0)\big) = P(X_i > m_0)(1 - P(X_i > m_0)) = 1/4$

Under $H_0: m = m_0$. Therefore,

$$\frac{S_n/n - 1/2}{\sqrt{1/4}/\sqrt{n}} = \frac{S_n - n/2}{\sqrt{n/4}} \sim N(0,1) \text{ for sufficiently large } n.$$

We can test $H_0: m = m_0$ using this fact.

Sign test : large sample approximation

Ex1 : revisited] A random sample with size 20 was observed:

(1.12 0.84 -0.71 -0.38 -4.05 -0.08 1.68 -0.49 0.17 3.51 -0.04 0.19 0.13 0.91 -0.11 0.74 9.33 0.35 -3.04 0.04).

We want to test

$$H_0: m = 0 \qquad vs \qquad H_1: m \neq 0$$

Answer] The number of observations greater than 0 is 12. Then, the observed test statistic by the large sample approximation is

$$\frac{12 - 20/2}{\sqrt{20/4}} = 0.89,$$

And therefore the p-value is given by

$$P(|Z| \geq 0.89) = 0.3735$$

where Z stands for the standard normal distribution.

Note that the exact p-value is

$$2 \cdot P(B \geq 12) = 0.5034$$

where $B \sim Binom\ (20,1/2)$.

Wilcoxon Signed Rank test : large sample approximation

The test statistic for the sign test is

$$W_n = \sum_{i=1}^{n} R_i^+ I(X_i - m_0 > 0).$$

This can be rewritten as

$$W_n = \sum_{i=1}^{n} i\, W_i\,,$$

where $W_i = I\big(i^{\text{th}}$ largest $|X_i - m_0|$ corresponds to some positive $X_i - m_0\big)$, and

(1) $E(W_n) = n(n+1)/4$

(2) $\mathrm{Var}(W_n) = n(n+1)(2n+1)/24.$

By a version of CLT, it can be shown that

$$\frac{W_n - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim\ N(0,1)\ \text{ for sufficiently large } n.$$

under $H_0\!: m = m_0$. Therefore, we can test $H_0\!: m = m_0$ using the standard normal distribution. The technical detail is out of scope.

Wilcoxon Signed Rank test : large sample approximation

Ex1 : revisited] A random sample with size 20 was observed:

(1.12 0.84 -0.71 -0.38 -4.05 -0.08 1.68 -0.49 0.17 3.51 -0.04 0.19 0.13 0.91 -0.11 0.74 9.33 0.35 -3.04 0.04).

We want to test

$$H_0 : m = 0 \qquad vs \qquad H_1 : m \neq 0$$

Answer] Recall that

$$w_n = 135.5.$$

Then, the observed test statistic by the large sample approximation is

$$\frac{135.5 - 20 \times 21/4}{\sqrt{20 \times 21 \times 41 \, /24}} = 1.14,$$

and therefore the p-value is

$$P(|Z| \geq 0.89) = 0.2543$$

where Z stands for the standard normal distribution.

Note that the exact p-value is 0.2627.