

Sign test: introduction

The Sign test is a nonparametric test in the sense that we do NOT make the assumption that the data were generated from a parametric distribution.

The Sign test allows us to test hypotheses on a location parameter, typically the median, but it could be another quantile of the distribution. We also do NOT assume that this distribution is continuous.

0

The bilateral test only exists when we do the test on the median. Unilateral tests could be done on other quantiles.

Test statistic (1/2)

Let us consider a Sign test, which has null hypothesis $H_0 : \text{med}(X) = \theta_0$ with $\theta_0 \in \mathbb{R}$, i.e. H_0 : 'The median of the data is θ_0 '. We observe a sample x_1, \dots, x_n . The Sign test statistic is

$$T(x_1, \dots, x_n) = \sum_{i=1}^n \mathbb{1}_{x_i > \theta_0} = \sum_{i=1}^n \mathbb{1}_{x_i - \theta_0 > 0}$$

where

$$\mathbb{1}_{x_i > \theta_0} = \begin{cases} 1, & \text{if } x_i > \theta_0. \\ 0, & \text{otherwise.} \end{cases}$$

Test statistic (2/2)

When none of the x_i are equal to θ_0 , then the test statistic can be rewritten as $T(x_1, \dots, x_n) = (S + n)/2$ where S is equal to

$$S = \sum_{i=1}^n \text{sign}(x_i - \theta_0)$$

with $\text{sign}(x_i - \theta_0) = 1$ if $x_i > \theta_0$ and -1 if $x_i < \theta_0$. Those two versions of the Sign test are equivalent.

Distribution of the test statistic

Under H_0 , the distribution of $\mathbb{1}_{x_i > \theta_0}$ is Bernoulli with parameter $p = 1 - F_X(\theta_0)$, where F_X is the CDF of X . It follows that the test statistic is Binomial, with $p = 1/2$ if we do test on the median. We have indeed

$$T(x_1, \dots, x_n) \sim B(n, p = 1/2)$$

So that we have

$$E[\mathbb{1}_{x_i > \theta_0}] = p = 1/2 \quad \text{and variance } \text{var}(\mathbb{1}_{x_i > \theta_0}) = p(1-p) = 1/4$$

$$\text{and } E[T(x_1, \dots, x_n)] = E[\sum_{i=1}^n \mathbb{1}_{x_i > \theta_0}] = np = n/2 \text{ and variance } \text{var}(T(x_1, \dots, x_n)) = np(1-p) = n/4.$$

Asymptotic distribution of the test statistic

In addition, from the CLT, it follows that

$$\frac{\sum_{i=1}^n x_i - E[X]}{\sqrt{n} \text{sd}(X)} \sim N(0, 1)$$

$$\sum_{i=1}^n x_i \sim N(nE[X], n\text{var}(X))$$

$$\frac{1}{n} \sum_{i=1}^n x_i \sim N(E[X], \frac{\text{var}(X)}{n})$$

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x_i > \theta_0} \sim N(1/2, 1/4n)$$

$$T(x_1, \dots, x_n) = \sum_{i=1}^n \mathbb{1}_{x_i > \theta_0} \sim N(n/2, n/4)$$

Asymptotic critical region

Let us in addition consider an alternative hypothesis $H_1 : \text{med}(X) \neq \theta_0$, that allows us to define a critical region R_α . If $T(x_1, \dots, x_n) \in R_\alpha$, H_0 is rejected at level α . Else, there is not enough evidence against H_0 at level α . The critical region is given by

$$R_\alpha = \{0, \dots, k\} \cup \{n - k, \dots, n\}$$

where k is the largest integer such that $P_{H_0}(T(x_1, \dots, x_n) \in R_\alpha) \leq \alpha$. Using the CLT, the large sample critical region is then becomes

$$R_\alpha = \left\{0, \dots, n/2 - q_{1-\frac{\alpha}{2}} \sqrt{n/4}\right\} \cup \left\{n/2 + q_{1-\frac{\alpha}{2}} \sqrt{n/4}, \dots, n\right\}$$

where $q_{1-\frac{\alpha}{2}} \approx 1.96$ if $\alpha = 0.05$.

Working example 1 (1/3)

Example 1: We test the hypothesis $H_0 : med(X) = 280$. We want to perform the test at level $\alpha = 5\%$ against the alternative hypothesis $H_1 : med(X) \neq 280$. Suppose that we observe the following data.

n	x_i	$ x_i - 280 $	$sign(x_i - 280)$
1	275	15	-
2	292	12	+
3	281	1	+
4	284	4	+
5	285	5	+
6	283	3	+
7	290	10	+
8	294	14	+
9	300	20	+
10	284	4	+

Working example 1 (2/3)

The Sign test statistic is therefore equal to $T(x_1, \dots, x_{10}) = (S + n)/2 = (9 + 10)/2 \approx 9$ or $\sum_{i=1}^n \mathbb{1}_{x_i - \theta_0 > 0} = 9$. If we choose $k = 1$, the critical region at level $\alpha = 5\%$ is then

$$R_{0.05} = \{0, \dots, 1\} \cup \{9, \dots, 10\}$$

Since $T(x_1, \dots, x_{10}) \in R_{0.05}$, we conclude that H_0 is rejected and that the median of the data is NOT equal to 280. We note that the true level of the test (probability of type I error) is equal to 2.15% and not 5%.

```
1 > 2*pbinom(q = 1, size = 10, prob = 1/2)
2 [1] 0.02148438
```


Working example 1 (3/3)

If we choose $k = 2$, the critical region becomes

$$R_{0.05} = \{0, \dots, 2\} \cup \{8, \dots, 10\},$$

and the level of the test is

```
> 2*pbinom(q = 2, size = 10, prob = 1/2)
[1] 0.109375
```

So we conclude that $k = 1$ is optimal in this case.

R code for working example 1

```
1 # data
2 data <- c(275,292,281,284,285,283,290,294,300,284)
3
4 # sign test statistic
5 dataminusmed0 <- data -280
6 S = (sum(sign(dataminusmed0)) + length(data) ) / 2
7 S # 9
8
9 # p-value
10 2*pbinom(q=1, size = length(data), prob = 0.5)
11 # [1] 0.02148438
12
13 # with package BSDA
14 library(BSDA)
15
16 SIGN.test(x = data, y = NULL, md = 180, alternative = "two.sided", conf.level =
    0.95)
17
18 data: data
19 s = 10, p-value = 0.001953
20 alternative hypothesis: true median is not equal to 180
21 95 percent confidence interval:
22 281.6489 293.3511
23 sample estimates:
24 median of x
25 284.5
26
27
28 Conf.Level L.E.pt U.E.pt
29 Lower Achieved CI 0.8906 283.0000 292.0000
30 Interpolated CI 0.9500 281.6489 293.3511
31 Upper Achieved CI 0.9785 281.0000 294.0000
```

Python code for working example 1

```
1 import numpy as np
2 from scipy.stats import binom
3 import statsmodels
4 from statsmodels.stats.descriptivestats import sign_test
5
6 # data
7 data = [275,292,281,284,285,283,290,294,300,284]
8
9 # We test: H0: median(data) = 280
10
11 # sign test statistic
12 dataminusmed0 = data - np.repeat(280, 10)
13 dataminusmed0
14
15 S = (np.sum(np.sign(dataminusmed0)) + len(data) ) / 2
16 S # 9.0
17
18 # p-value
19 2*binom.pmf(1, n = 10, p = 0.5)
20 # 0.019531250000000003
21
22 # p-value with method sign_test() from statsmodels
23 sign_test(data, mu0=180)
24 # (5.0, 0.001953125)
```

Working example 2

Example 2: We consider a sample of $n = 64$ observations. Out of these 64 values, 20 are larger than some median θ_0 . What is the critical region R_α and the p-value of a bilateral test at $\alpha = 5\%$?

The critical region at level $\alpha = 5\%$ is then

$$\begin{aligned} R_{0.05} &= \left\{ 0, \dots, 64/2 - q_{1-\frac{0.05}{2}} \sqrt{64/4} \right\} \cup \\ &\quad \left\{ 64/2 + q_{1-\frac{0.05}{2}} \sqrt{64/4}, \dots, 64 \right\} \\ &= \left\{ 0, \dots, 24 \right\} \cup \left\{ 40, \dots, 64 \right\}, \end{aligned}$$

which is centered on $n/2 = 64/2 = 32$. Since $T(x_1, \dots, x_n) \in R_\alpha$, we reject H_0 at 0.05 level of significance. The p-value is $2 * P(T \leq$

$$20) \quad \Leftrightarrow \quad 2 * P\left(\frac{T-32}{4} \leq \frac{20-32}{4}\right) \quad \Leftrightarrow \quad 2 * P(z \leq -3) \quad \Leftrightarrow$$

$$2 * \Phi(-3) = 2 * 0.01 = 0.02$$

References

Bagdonavičius V., Kruopis J., Nikulin M. S., Non-parametric Tests for Complete Data (2011), Wiley, ISBN 978-1-84821-269-5 (hard-back)

<https://www.rdocumentation.org/packages/BSDA/versions/1.2.1/topics/SIGN.test>

The R Project for Statistical Computing:

<https://www.r-project.org/>

Python:

<https://www.python.org/>

course notes