STA 104 Applied Nonparametric Statistics

Chapter 2: One-Sample Methods for Location Problem

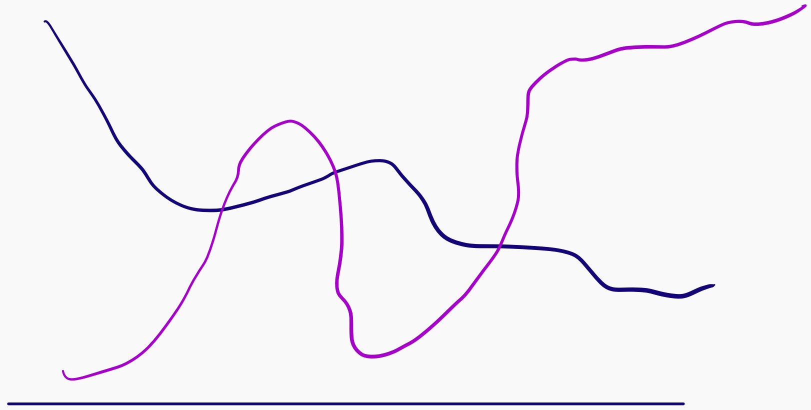
Xiner Zhou

Department of Statistics, University of California, Davis

Table of contents

Signed Test

Setting



Suppose we have a random sample $x_1 \dots x_n$

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

Hypothesis

Two-Sided Test:

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

One-Sided Upper-Tail Test:

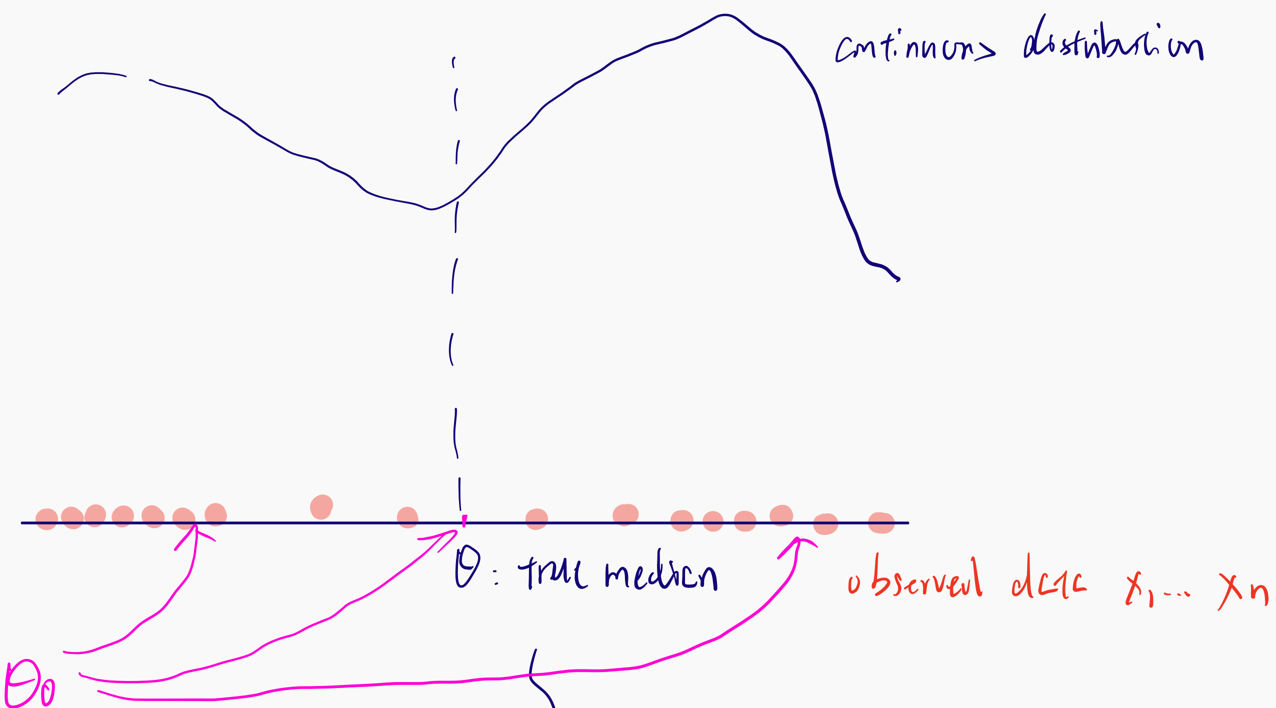
$$H_0 : \theta = \theta_0 \text{ versus } H_a : \theta > \theta_0$$

One-Sided Lower-Tail Test:

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

Motivation

- Centering: subtract θ_0 from each observation $x_1, \ldots, x_n$ to form a modified sample $x_1' = x_1 - \theta_0, \ldots, x_n' = x_n - \theta_0$
- Define indicator variables $\psi_i, i = 1, \ldots, n$ where $\psi_i = \begin{cases} 1, & \text{if } X_i > \theta_0 \\ 0, & \text{if } X_i < \theta_0 \end{cases}$
- Let $B = \sum_{i=1}^{n} \psi_i$ denote the number of X_i 's out of n that fall above the hypothesized median θ_0.



$P(X \geq \theta) = \frac{1}{2}$: probability that each sample is greater than true median is $\frac{1}{2}$.

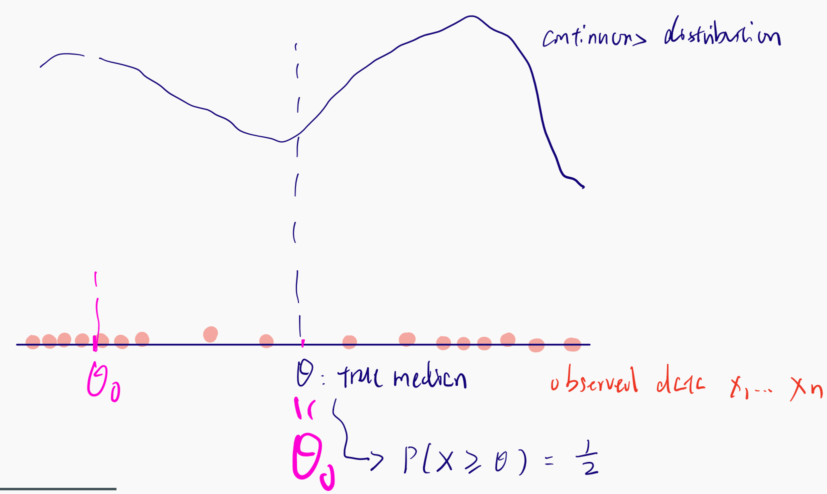
Motivation (upper tail hypothesis $H_0: \theta = \theta_0$ vs $H_a: \theta > \theta_0$)

If H_0 is true, then each X_i has probability .5 of falling above θ_0, so B has a binomial distribution with probability $p = .5$.

If the true median is greater than θ_0, then B has a binomial distribution with probability $p > .5$.

Thus, we can decide between H_0 and H_a based on the value of B; that is, we can test [1]

$$\underbrace{H_0: \theta = \theta_0 \text{ vs } H_a: \theta > \theta_0}_{\text{original one-sample median problem}} \iff \underbrace{H_0 : p = .5, H_a : p > .5}_{\text{Binomial problem}}$$



[1] The test procedures based on the sign statistic B are actually special cases of the general binomial test . The sign test procedures are simply binomial procedures, with "success" corresponding to a positive centered observation, "failure" corresponding to a negative centered observation, and $p = P(\text{"success"}) = P(X_i > \theta_0)$ assuming the value $p_0 = \frac{1}{2}$ when the null hypothesis $H_0 : \theta = \theta_0$ is true.

Procedure

a. One-Sided Upper-Tail Test.

To test

$$H_0 : \theta = \theta_0$$

versus

$$H_1 : \theta > \theta_0$$

at the α level of significance,

Reject H_0 if $B \geq b_\alpha$; otherwise do not reject, where the constant b_α is chosen to make the type I error probability equal to α. The number b_α is the upper α percentile point of the binomial distribution with sample size n and success probability p_0.

The normal approximation:

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

Procedure

b. One-Sided Lower-Tail Test.

To test

$$H_0 : \theta = \theta_0$$

versus

$$H_{\cancel{0}a} : \theta < \theta_0$$

at the α level of significance, Reject H_0 if $B \leq b_{1-\alpha}$; otherwise do not reject.

The normal approximation:

Reject H_0 if $B^* \leq -z_\alpha$; otherwise do not reject.

Procedure

c. Two-Sided Test.

To test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta \neq \theta_0$$

at the α level of significance, Reject H_0 if $B \geq b_{\alpha/2}$ or $B \leq b_{1-\alpha/2}$; otherwise do not reject

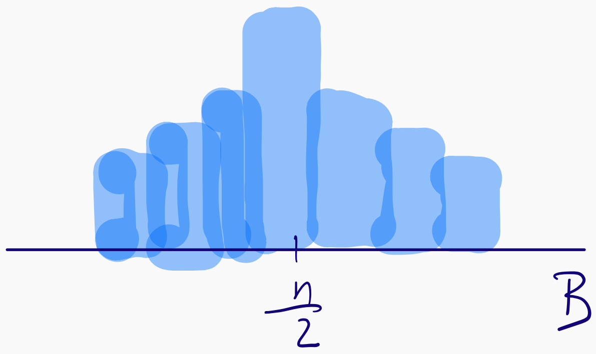
The normal approximation:

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

An estimator associated with the signed statistics (Hodges-Lehmann)

The null distribution of the statistic $B = \sum_{i=1}^{n} \psi_i \sim Bin(n, \frac{1}{2})$ is around its mean, $n/2$.

A natural estimator of θ is the amount that should be subtracted from each X_i so that the value of B, when applied to the shifted sample $X_1 - \hat{\theta}, \ldots, X_n - \hat{\theta}$, is as close to the center $n/2$ as possible.



$$B = \sum_{i=1}^{n} \psi_i \approx n/2$$
$$\Rightarrow \frac{\sum_{i=1}^{n} \psi_i}{n} \approx 1/2$$

i.e. total observations above the true population median is the sample median.

Intuitively, we estimate θ by the amount that the X sample should be shifted in order that $X_1 - \hat{\theta}, \ldots, X_n - \hat{\theta}$ appears (when "viewed" by the sign statistic B) as a sample from a population with median 0 .

Procedure

Estimate

The estimator of θ associated with the sign statistic

$$\hat{\theta} = \text{median}\left\{X_i, 1 \leq i \leq n\right\}.$$

Thus,

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

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

  the value that occupies position $k + 1$ in the list of the ordered X_i values.
- If n is even, say $n = 2k$, then $k = n/2$ and

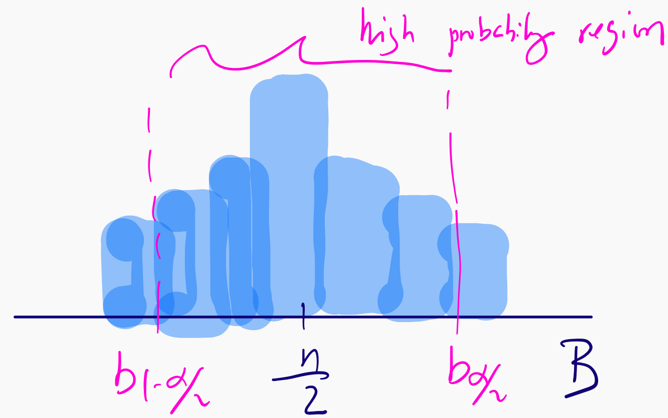
  $$\hat{\theta} = \frac{X^{(k)} + X^{(k+1)}}{2};$$

  that is, when n is even, $\tilde{\theta}$ is the average of the two X_i values that occupy positions k and $k + 1$ in the ordered list of the n data values.

Confidence interval based on the signed test

The true population median θ_0 is the value such that the number of observations above it is the signed statisitcs B which should be centered at $\frac{n}{2}$ with some natural variation.

So we use the natural variation of T^+ reverse engineer the most probable region of θ_0 .



$$\underbrace{P(b_{1-\alpha/2} \le B \le b_{\alpha/2}}_{\text{number of observations above the median should be between } b_{1-\alpha/2} \text{ and } b_{\alpha/2}}) = 1-\alpha$$

$\Rightarrow$ median should be between the $b_{1-\alpha/2}$-th and $b_{\alpha/2}$-th ordered observations.

Procedure

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

For a symmetric two-sided confidence interval for θ, with confidence coefficient $1-\alpha$, first obtain the upper $(\alpha/2)$ nd percentile point $\boxed{b_{\alpha/2}}$ of the null distribution of $\boxed{B \sim Bin(n, 1/2)}$

$$\cancel{b_{1-\alpha/2}} = n + 1 - b_{\alpha/2}. \quad \leftarrow \text{lower limit.}$$

The $100(1-\alpha)\%$ confidence interval $(\theta_{\mathrm{L}}, \theta_{\mathrm{U}})$ for θ that is associated with sign test

$$\theta_{\mathrm{L}} = X^{\left(b_{1-\alpha/2}\right)}, \; \theta_{\mathrm{U}} = X^{\left(b_{\alpha/2}\right)}$$

where $X^{(1)} \leq \cdots \leq X^{(n)}$ are the ordered sample observations; that is, θ_L is the sample observation that occupies position $b_{1-\alpha/2}$ in the list of ordered sample data. The upper end point θ_U is the sample observation that occupies position $b_{\alpha/2}$ in this ordered list.

Then we have

$$P_{\theta}\left(\theta_{\mathrm{L}} < \theta < \theta_{\mathrm{U}}\right) = 1 - \alpha \text{ for all } \theta.$$

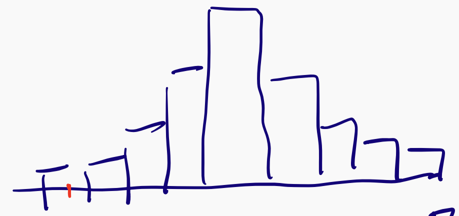
Example: The Mariner and the Pioneer Spacecraft Data

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 $\theta \neq 81.3035$.

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

$\alpha = 0.05$

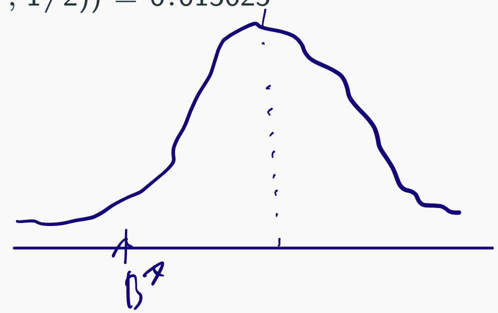


$B \sim Bin(7, \frac{1}{2})$

Exact test:

$$B = 0$$

$$p - value = 2P(B \leq 0 | B \sim Bin(7, 1/2)) = 0.015625$$



Large-sample approximation:

$$B^* = \frac{0 - \left(\frac{7}{2}\right)}{\left(\frac{7}{4}\right)^{1/2}} = -2.645751$$

$Var(B)$

$$p - value = 2P(Z < -2.645751) = 0.008150979$$

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 $\theta = 81.3035$.

Compare with Wilcoxon, no qualitative difference (minor quantitative difference)

An estimate for median:

The ordered Z observations are $Z^{(1)} \leq \cdots \leq Z^{(7)}$: $\cancel{81.299781.300181.300581.300681.301181.301581.3021}$

$$\hat{\theta} = Z^{(4)} = 81.3006$$

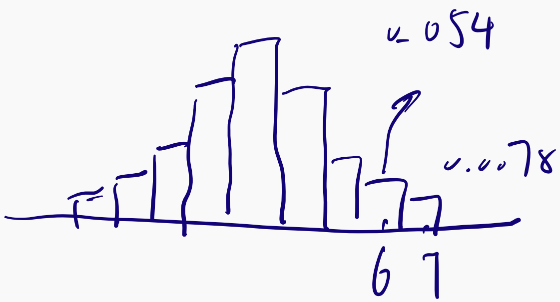
Confidence interval for median:

With $n = 7$ and $\alpha = .05$, the null distribution of B:

exact null distribution of $B \sim \text{Binomial}(n, p)$

$= \text{Bin}(7, \frac{1}{2})$

```
> dbinom(x=seq(0,7,by=1), size=7, prob=0.5)
[1] 0.0078125 0.0546875 0.1640625 0.2734375 0.2734375 0.1640625 0.0546875 0.0078125
```



not use this notation

$$b_{\alpha/2} = 7,\ \cancel{b_{1-\alpha/2}} = \underbrace{8}_{n+1=7+1} - 7 = 1$$

$b_{0.025}$

$$\theta_{\mathrm{L}} = Z^{(1)} = 81.2997 \text{ and } \theta_{\mathrm{U}} = Z^{(7)} = 81.3021$$

so that our 95% confidence interval for θ is

$$(\theta_{\mathrm{L}}, \theta_{\mathrm{U}}) = (81.2997, 81.3021)$$

corresponds to $\underbrace{\text{Upper Achieved CI}}_{\text{conservative}}$.

conservative : actual confidence level is > 95%

```
> library(BSDA)
> SIGN.test(c(81.3001,81.3015,81.3006,81.3011,81.2997,81.3005,81.3021), md=14)

        One-sample Sign-Test

data:  c(81.3001, 81.3015, 81.3006, 81.3011, 81.2997, 81.3005, 81.3021)
s = 7, p-value = 0.01563
alternative hypothesis: true median is not equal to 14
95 percent confidence interval:
 81.29983 81.30191
sample estimates:
median of x
    81.3006

Achieved and Interpolated Confidence Intervals:

                  Conf.Level  L.E.pt  U.E.pt
Lower Achieved CI     0.8750 81.3001 81.3015
Interpolated CI       0.9500 81.2998 81.3019
Upper Achieved CI     0.9844 81.2997 81.3021
```

← exact p-value

actual confidence > 95%