ST 516: Foundations of Data Analytics

Sign Test

Inference for Population Median

So far, the inference procedures (point estimates, confidence intervals, and hypothesis tests) that we have considered have all focused on the population mean parameter μ.

What if we are interested in a different population parameter? Here we will present inference for the population *median*.

Recall that the population median of a variable X is the value M such that 50% of the population values of X are less than M, and 50% are greater than M. The median is another measure of the *location* of the population distribution: it tells us about a ‘typical’ value of X.

When, Why Use the Median?

You may encounter arguments that the median should be used when the data are heavily skewed (asymmetric), or when there are severe outliers.

Neither of these recommendations are entirely complete or accurate: your choice of which parameter to use should depend on the scientific context.

Means are useful when you want to capture a *total* population tendency. Medians are useful when you are interested in where the boundary between the upper half and lower half of the values in the population lies.

When, Why Use the Median?

For instance, suppose you are considering a vacation package. Would you find it more useful to know the *mean* cost per day, or the *median* cost per day?

When, Why Use the Median?

For instance, suppose you are considering a vacation package. Would you find it more useful to know the *mean* cost per day, or the *median* cost per day?

What if the daily costs for a representative seven day trip are:

$50, $60, $80, $40, $90, $70, $800?

The mean daily cost is $170 per day; the median is $70. Which summary gives you a better idea of how much money you will spend on this trip?

When, Why Use the Median?

As another example, suppose you are curious about class sizes at a local university. Would you find it more useful to know the *mean* size of a class, or the *median* size of a class?

What if a sample of 9 classes has the following class sizes (number of students):

$$7, \quad 12, \quad 16, \quad 5, \quad 10, \quad 19, \quad 14, \quad 8, \quad 233$$

The mean number of students per class is 36; the median is 12. Which summary gives you a better idea of how many students might be in a typical class?

Inference for Population Median

Just like we did for the population mean parameter, we can estimate the population median M using the sample median m:

$$m = \text{Middle value of the sample}$$

This estimate is *unbiased*: The expected value (mean) of the sampling distribution of the sample median is equal to the population median. In notation,

$$\mathsf{E}(m) = M$$

In other words, across repeat experiments, the average value of the sample median will be equal to the population median.

Inference for Population Median

Recall that the Central Limit Theorem gives us an approximate sampling distribution for the sampling distribution of the sample mean. However, the Central Limit Theorem does not apply directly to the the sampling distribution of the sample median. The sampling distribution of the sample median is a bit more complicated to express, even in large sample sizes.

We would still like to construct confidence intervals and perform hypothesis tests for the population median, but we cannot directly apply the exact same approaches we did for the population mean.

Hypothesis Test for Population Median

Instead, we can use the following approach to test the null hypothesis that the population median is M_0 for some specified value M_0. ($H_0 : M = M_0$ vs. $H_A : M \neq M_0$):

- If M_0 *is* the true population median, we expect approximately 50% of our sample to be smaller than M_0.
- We can think of each observation as a 'trial', which is a 'success' if that observation is smaller than M_0 and a 'failure' otherwise. If our null hypothesis $H_0 : M = M_0$ is *true*, then each trial has 50% chance of 'success': $\pi = P(X_i \leq M_0) = 0.5$.
- We can count how many 'successes' we have in our n 'trials'. If the null hypothesis $H_0 : M = M_0$ is true, this count will follow a Binomial distribution with n trials and success probability $\pi = 0.5$. We can test that the success probability is indeed $\pi = 0.5$ using the Binomial proportion test.

Sign Test

Formally, what we just described is a test called the **Sign Test**, which tests the hypothesis $H_0: M = M_0$ vs. $H_A: M \neq M_0$. Here is the more formal description of the sign test:

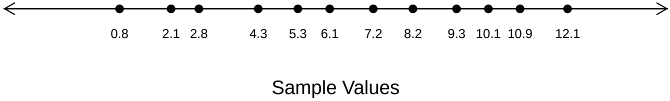
- Let

  $$Y_i = \begin{cases} 1 & \text{if } X_i \leq M_0 \\ 0 & \text{otherwise} \end{cases}$$

  so Y_i are independent binary random variables with success probability $\pi = P(Y_i = 1) = P(X_i \leq M_0)$.

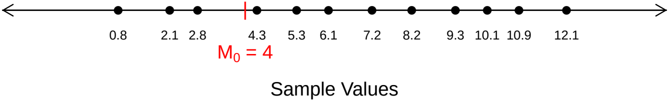
- Use the test of a Binomial proportion on the Y_is to test $H_0: \pi = 0.5$ vs. $H_A: \pi \neq 0.5$.

Sign Test Example



Consider the example data above.

Sign Test Example

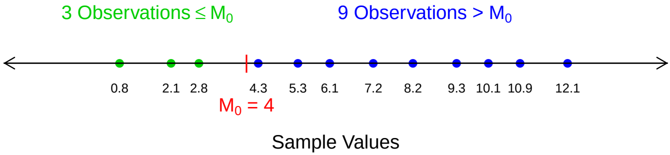


Consider the example data above.

Suppose we want to perform a level $\alpha = 0.05$ test of the null hypothesis that the population median is $M_0 = 4$:

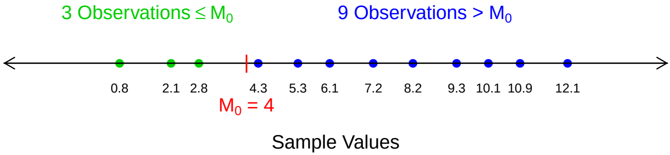
$$H_0 : M = 4$$

Sign Test Example



First we count how many observations are less than or equal to $M_0 = 4$.

Sign Test Example

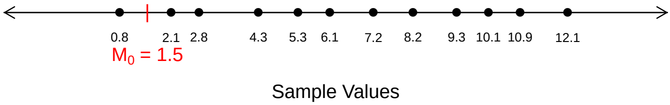


Then we perform a binomial proportion test to test that $\pi = P(X_i \leq M_0)$ is equal to $\pi_0 = 0.5$, using the observed sample proportion $\hat{\pi}_{M_0} = \dfrac{3}{12} = 0.25$:

$$Z = \frac{\hat{\pi}_{M_0} - \pi_0}{\sqrt{(\pi_0(1-\pi_0))/n}} = \frac{0.25 - 0.5}{\sqrt{(0.5(0.5))/12}} = -1.732$$

Since $|Z| = 1.732 < z_{\alpha/2} = 1.96$, we *fail to reject* the null hypothesis $H_0 : M = 4$.

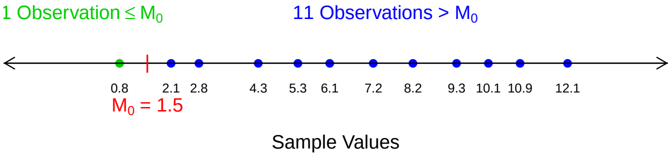
Sign Test Example



Now suppose instead we want to perform a level $\alpha = 0.05$ test of the null hypothesis that the population median is $M_0 = 1.5$:

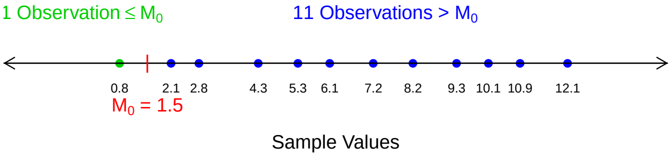
$$H_0 : M = 1.5$$

Sign Test Example



First we count how many observations are less than or equal to $M_0 = 1.5$.

Sign Test Example



Then we perform a binomial proportion test to test that $\pi = P(X_i \leq M_0)$ is equal to $\pi_0 = 0.5$, using the observed sample proportion $\hat{\pi}_{M_0} = \dfrac{1}{12} = 0.083$:

$$Z = \frac{\hat{\pi}_{M_0} - \pi_0}{\sqrt{(\pi_0(1-\pi_0))/n}} = \frac{0.083 - 0.5}{\sqrt{(0.5(0.5))/12}} = -2.887$$

Since $|Z| = 2.887 > z_{\alpha/2} = 1.96$, we *reject* the null hypothesis $H_0 : M = 1.5$.

Confidence Interval for Population Median

To construct confidence intervals for population medians, we recall the duality between confidence intervals and hypothesis tests:

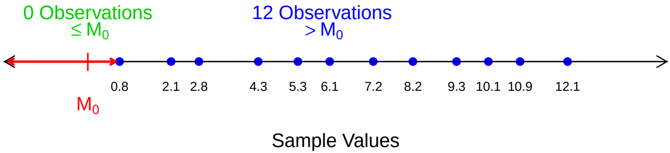
A $(1-\alpha)100\%$ confidence interval for a parameter θ is the set of all values θ_0 for which a level α test of $H_0: \theta = \theta_0$ vs. $H_A: \theta \neq \theta_0$ would not reject the null hypothesis.

Using this duality, we can use the sign test to construct a confidence interval for the population median: **a $(1-\alpha)100\%$ confidence interval for M is just the set of values M_0 that would not be rejected at level α.**

Note that we want an interval for M, not for $\pi = P(X_i \leq M_0)$, so we can't just use the binomial proportion confidence interval!

Confidence Interval Demonstration

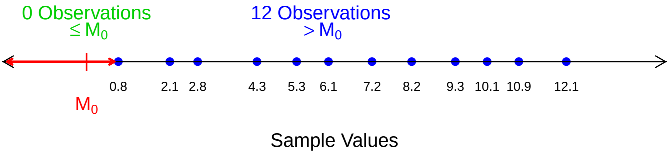
Using our example data, let's consider what would happen if we tested different values of M_0.



First, we consider M_0 in the range $(-\infty, 0.8)$. For instance, $H_0 : M = 0$.

Confidence Interval Demonstration

Using our example data, let's consider what would happen if we tested different values of M_0.

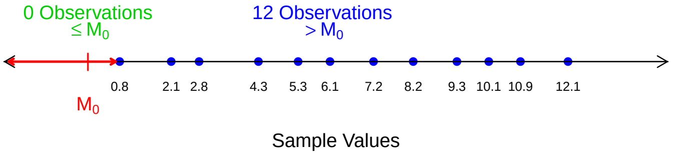


The z-statistic based on the observed proportion $\hat{\pi}_{M_0} = \dfrac{0}{12}$ for testing $H_0 : \pi = 0.5$ is

$$Z = \frac{0 - 0.5}{\sqrt{(0.5(0.5))/12}} = -3.464$$

Confidence Interval Demonstration

Using our example data, let's consider what would happen if we tested different values of M_0.

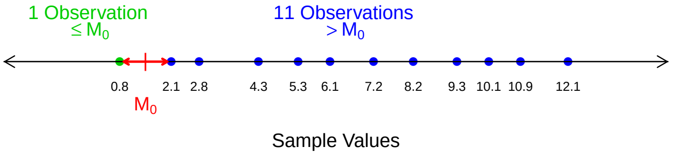


Since $|Z| = 3.464 > z_{0.025} = 1.96$, we **REJECT** $H_0 : M = 0$ (or any other value in the range $(-\infty, 0.8)$.

Therefore, $M = 0$ **IS NOT** in the 95% confidence interval for M.

Confidence Interval Demonstration

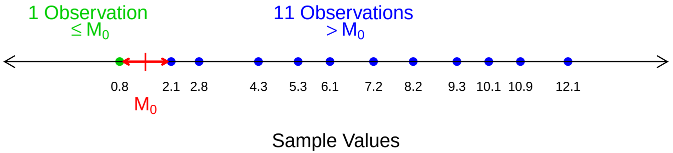
Using our example data, let's consider what would happen if we tested different values of M_0.



Now let's do the same thing for other ranges of M_0.

Confidence Interval Demonstration

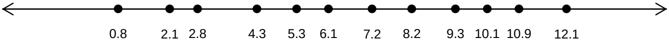
Using our example data, let's consider what would happen if we tested different values of M_0.



M_0 in the range $(0.8, 2.1)$: For instance, $H_0 : M = 1.5$.

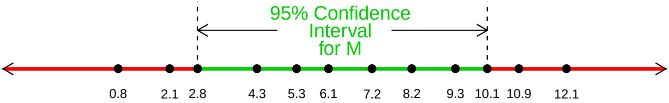
M_0	$\hat{\pi}_{M_0}$	Z	Decision	M_0 In 95% CI?
1.5	$\frac{1}{12}$	$\frac{\frac{1}{12} - 0.5}{\sqrt{(0.5(0.5))/12}}$ = -2.887	**REJECT** $H_0 : M = 1.5$	**NO**

Confidence Interval Demonstration



M_0 Range	$\hat{\pi}_{M_0}$	Z	Decision	M_0 In 95% CI?
$(\infty, 0.8)$	$0/12$	-3.464	REJECT	NO
$(0.8, 2.1)$	$1/12$	-2.887	REJECT	NO
$(2.1, 2.8)$	$2/12$	-2.309	REJECT	NO
$(2.8, 4.3)$	$3/12$	-1.732	Do not reject	YES
$(4.3, 5.3)$	$4/12$	-1.155	Do not reject	YES
$(5.3, 6.1)$	$5/12$	-0.577	Do not reject	YES
$(6.1, 7.2)$	$6/12$	0.000	Do not reject	YES
$(7.2, 8.2)$	$7/12$	0.577	Do not reject	YES
$(8.2, 9.3)$	$8/12$	1.155	Do not reject	YES
$(9.3, 10.1)$	$9/12$	1.732	Do not reject	YES
$(10.1, 10.9)$	$10/12$	2.309	REJECT	NO
$(10.9, 12.1)$	$11/12$	2.887	REJECT	NO
$(12.1, \infty)$	$12/12$	3.464	REJECT	NO

Confidence Interval Demonstration



M_0 Range	$\hat{\pi}_{M_0}$	Z	Decision	M_0 In 95% CI?
$(\infty, 0.8)$	$0/12$	-3.464	REJECT	NO
$(0.8, 2.1)$	$1/12$	-2.887	REJECT	NO
$(2.1, 2.8)$	$2/12$	-2.309	REJECT	NO
$(2.8, 4.3)$	$3/12$	-1.732	Do not reject	YES
$(4.3, 5.3)$	$4/12$	-1.155	Do not reject	YES
$(5.3, 6.1)$	$5/12$	-0.577	Do not reject	YES
$(6.1, 7.2)$	$6/12$	0.000	Do not reject	YES
$(7.2, 8.2)$	$7/12$	0.577	Do not reject	YES
$(8.2, 9.3)$	$8/12$	1.155	Do not reject	YES
$(9.3, 10.1)$	$9/12$	1.732	Do not reject	YES
$(10.1, 10.9)$	$10/12$	2.309	REJECT	NO
$(10.9, 12.1)$	$11/12$	2.887	REJECT	NO
$(12.1, \infty)$	$12/12$	3.464	REJECT	NO

We see that values of M_0 between 2.8 and 10.1 are in the 95% confidence interval for M, so our 95% confidence interval is $(2.8, 10.1)$.

Confidence Interval for Population Median

Using this same approach, we can work out more generally what the confidence interval limits should be for any given data-set/sample size by solving for which values of M_0 would not be rejected.

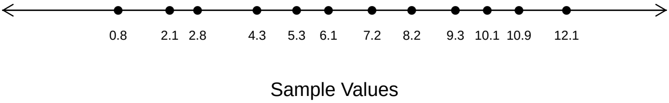
This gives the following *approximate* $(1-\alpha)100\%$ confidence interval for the population median:

$$\left(\left(\frac{n}{2}-\frac{z_{\alpha/2}\sqrt{n}}{2}\right)\text{th Smallest Observation},\right.$$
$$\left.\left(\frac{n}{2}+\frac{z_{\alpha/2}\sqrt{n}}{2}+1\right)\text{th Smallest Observation}\right)$$

Note that often the values $\frac{n}{2}-\frac{z_{\alpha/2}\sqrt{n}}{2}$ and $\frac{n}{2}+\frac{z_{\alpha/2}\sqrt{n}}{2}+1$ will not be integers, in which case we round to the nearest integer.

Confidence Interval for Population Median: Example

Consider the same example data set:



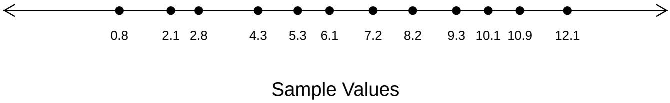
If we want a $(1-\alpha)100\% = 95\%$ confidence interval for the population median, we calculate

$$\frac{n}{2} - \frac{z_{0.025}\sqrt{n}}{2} = \frac{12}{2} - \frac{1.96\sqrt{12}}{2} = 2.605$$

$$\frac{n}{2} + \frac{z_{0.025}\sqrt{n}}{2} + 1 = \frac{12}{2} + \frac{1.96\sqrt{12}}{2} + 1 = 10.395$$

Confidence Interval for Population Median: Example

Consider the same example data set:



Rounding these numbers to integers, we find that the 95% confidence interval for the population median is from the 3rd smallest to the 10th smallest values in our sample.

- The 3rd smallest value is 2.8
- The 10th smallest value is 10.1

So the 95% confidence interval for the population median is

$$(2.8, 10.1)$$

just as we found earlier.

Confidence Interval for Population Median

Some comments on confidence intervals for the population median:

- Note that in different textbooks and literature, you may find slightly different versions of these confidence limits.
- All versions will be approximately equivalent when the sample size n is large, but may give different confidence intervals for small sample sizes.
- All of confidence intervals that are based on the Normal approximation to the Binomial distribution are *approximate*: they will not have exactly the desired coverage probability for small sample sizes n.