Chapter 7: One sample tests

(Ott & Longnecker Sections: 10.2)

Part 3

https://dzwang91.github.io/stat324/

WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON

Outline

1. Sign test

2. Tests for population proportion

Sign test

Definition

For a random variable X, population median is defined as a parameter m such that

$$P(X \leq m) = \frac{1}{2}$$

Sign test

Definition

For a random variable X, population median is defined as a parameter m such that

$$P(X \leq m) = \frac{1}{2}$$

- Collect an i.i.d. sample $X_1, \ldots, X_n$ from some population with population median m.
- **Step 1**:
  1. Two-tailed test:
  $$H_0 : m = m_0 \text{ v.s. } m \neq m_0$$
  2. Left-tailed test:
  $$H_0 : m = m_0 \text{ v.s. } m < m_0$$
  3. Right-tailed test:
  $$H_0 : m = m_0 \text{ v.s. } m > m_0$$

Sign test

- **Step 2**: delete data points which are equal to m_0. Count remaining sample size n^*.

Sign test

- **Step 2**: delete data points which are equal to m_0. Count remaining sample size n^*.
- **Step 3**: choose test statistic

  $B =$ number of data values greater than m_0 in the remaining sample.

Sign test

- **Step 2**: delete data points which are equal to m_0. Count remaining sample size n^*.
- **Step 3**: choose test statistic

  $B =$ number of data values greater than m_0 in the remaining sample.

- Under $H_0 : m = m_0$, $B \sim \text{Binomial}(n^*, 0.5)$.

Sign test

- **Step 2**: delete data points which are equal to m_0. Count remaining sample size n^*.
- **Step 3**: choose test statistic

  $B =$ number of data values greater than m_0 in the remaining sample.

- Under $H_0 : m = m_0$, $B \sim \text{Binomial}(n^*, 0.5)$.
- **Step 4**: calculate p-value. Let b be the observation of B.
  - $H_A : m > m_0$:

    $$\text{p-value} = P(B \geq b \mid B \sim \text{Binomial}(n^*, 0.5))$$

  - $H_A : m < m_0$:

    $$\text{p-value} = P(B \leq b \mid B \sim \text{Binomial}(n^*, 0.5))$$

  - $H_A : m \neq m_0$:

    $$\text{p-value} = 2 \times \min\{P(B \geq b), P(B \leq b)\}.$$

Example

The sanitation department in a large city is considering separating recyclable material out of the trash to save on landfill space and make money selling the recyclables. Based on data from other cities, it is determined that if at least half of the households in the city have more than 4.6 lbs of reclaimable recyclable material, then the separation will be profitable for the city. A random sample of 11 households yields the following data on pounds of recyclable material found in the trash:

14.2, 5.3, 2.9, 4.2, 1.2, 4.3, 1.1, 2.6, 6.7, 7.8, 25.9

Note that the median of the sample is 4.3 lbs.

Example continued

- $H_0 : m = 4.6$ vs. $H_A : m > 4.6$.
- test statistic: B=the number of observations in the sample that are greater than 4.6
- Under H_0, $B \sim Bin(11, 0.5)$.
- Observation of B:

  14.2, 5.3, 2.9, 4.2, 1.2, 4.3, 1.1, 2.6, 6.7, 7.8, 25.9

  so $B_{obs} = 5$.
- p-value$= P(B \geq 5) = 0.726$.
- Set $\alpha = 0.05$, then there is no sufficient evidence to reject the null.
- There is no strong evidence that separation of the recyclables would be profitable in this city.

Tests for symmetric distributions

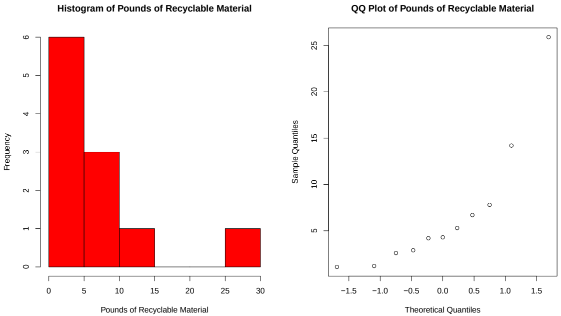
- For symmetric distributions, population median is equal to population mean, so T test and sign test can be used to test both of population mean and population median.

Tests for symmetric distributions

- For symmetric distributions, population median is equal to population mean, so T test and sign test can be used to test both of population mean and population median.
- Comparison between T test and sign test: T test requires the population distribution is normal, but sign test doesn't

Back to example

- Histogram and QQ plot:



The histogram looks neither normal nor symmetric, and the QQ plot also does not support normality. Therefore, we can not use T test.

Outline

1. Sign test

2. Tests for population proportion

Tests for population proportion

- Let π be the population proportion. We want to test

$$\text{(right-tailed test)} \quad H_0 : \pi = \pi_0, \; v.s. \; H_A : \pi > \pi_0;$$

or

$$\text{(left-tailed test)} \quad H_0 : \pi = \pi_0, \; v.s. \; H_A : \pi < \pi_0;$$

or

$$\text{(two-tailed test)} \quad H_0 : \pi = \pi_0, \; v.s. \; H_A : \pi \neq \pi_0.$$

Tests for population proportion

- Let π be the population proportion. We want to test

$$\text{(right-tailed test)} \quad H_0 : \pi = \pi_0, \ \ v.s. \ H_A : \pi > \pi_0;$$

or

$$\text{(left-tailed test)} \quad H_0 : \pi = \pi_0, \ \ v.s. \ H_A : \pi < \pi_0;$$

or

$$\text{(two-tailed test)} \quad H_0 : \pi = \pi_0, \ \ v.s. \ H_A : \pi \neq \pi_0.$$

- We have data $X_1, ..., X_n \sim Ber(\pi)$.

Tests for population proportion

- Let π be the population proportion. We want to test

$$\text{(right-tailed test)} \quad H_0 : \pi = \pi_0, \;\; v.s. \;\; H_A : \pi > \pi_0;$$

  or

$$\text{(left-tailed test)} \quad H_0 : \pi = \pi_0, \;\; v.s. \;\; H_A : \pi < \pi_0;$$

  or

$$\text{(two-tailed test)} \quad H_0 : \pi = \pi_0, \;\; v.s. \;\; H_A : \pi \neq \pi_0.$$

- We have data $X_1, ..., X_n \sim Ber(\pi)$.
- Under $H_0 : \pi = \pi_0$, the sample proportion $P = \frac{\sum_{i=1}^n X_i}{n}$ has

$$\mathbb{E}(P) = \pi_0$$

  and

$$Var(P) = \frac{\pi_0(1 - \pi_0)}{n}.$$

Tests for population proportion

- When $n\pi_0 > 5$ and $n(1-\pi_0) > 5$, the CLT holds:

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$

is approximately $N(0,1)$.

Tests for population proportion

- When $n\pi_0 > 5$ and $n(1-\pi_0) > 5$, the CLT holds:

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$

  is approximately $N(0, 1)$.

- **Test statistic**:

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$

  This is Z test!

- **P-value**: Let P_{obs} be the observed sample proportion.
  - $H_A : \pi > \pi_0$: calculate $z_{obs} = \frac{P_{obs} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ and calculate $P(Z > z_{obs})$.
  - $H_A : \pi < \pi_0$: calculate $P(Z < z_{obs})$.
  - $H_A : \pi \neq \pi_0$: calculate $2 * P(Z > |z_{obs}|)$.

Example

An accounting firm has a large list of clients, and each client has a file with information about that client. The firm has noticed errors in some of these files, and has decided that it would be worthwhile to know the proportion of files that contain an error. Call the population proportion of files in error π. It was decided to take a simple random sample of size $n = 50$, and use the results of the sample to estimate π. Each selected file was thoroughly reviewed, and classified as either containing an error (call this 1), or not (call this 0). The results are as follows:

Files with an error: 10; Files without any errors: 40.

Example continued

The company CEO decides that if π is greater than 0.1, then it will be worthwhile to review and fix every file. Therefore we wish to test:

Example continued

The company CEO decides that if π is greater than 0.1, then it will be worthwhile to review and fix every file. Therefore we wish to test:

$$H_0 : \pi = 0.1$$
$$H_A : \pi > 0.1.$$

Example continued

The company CEO decides that if π is greater than 0.1, then it will be worthwhile to review and fix every file. Therefore we wish to test:

$$H_0 : \pi = 0.1$$
$$H_A : \pi > 0.1.$$

- Under H_0, $0.1(50) = 5$ and $0.9(50) = 45 > 5$, so we should be able to use the CLT.
- Our observed statistic is $z_{obs} = \frac{0.2-0.1}{\sqrt{\frac{0.1(1-0.1)}{50}}} = 2.357$, so the p-value is $P(Z > 2.357) = 0.009$.
- At the 5% level ($\alpha = 0.05$), we would reject the null, and conclude that too high of a proportion of files are in error. All files should be checked and fixed.

What's the next?

We'll discuss tests for two independent samples in the next lecture.