Class Notes: STAT 501
Nonparametrics & Log-Linear Models
Sign test and Wilcoxon signed rank test

Da Kuang
University of Pennsylvania

Contents

1 Motivation of Nonparametric Methods

Traditional parametric testing methods are based on the assumption that data are generated by well-known distributions, characterized by one or more unknown parameters.

The critical values or alternatively the p-values is computed according to the distribution of the test statistic under the null hypothesis, which can be derived from the assumptions related to the assumed underlying distribution of data.

But the above assumption is not always true. For example, if the data is semi-continuous and also zero inflated, then any assumed distribution is not true. We can still apply the parametric testing method on the data but the result is probably not reasonable.

When the parametric does not hold, nonparametric methods are the valid solutions. Nonparametric methods require relatively mild assumptions regarding the underlying populations from which the data are obtained.

Note that when the parametric assumptions hold, the nonparametric methods are only slightly less powerful than the parametric methods.

2 Sign Test

Sign test is like the nonparametric version of the paired or one sample t-test. The primary interest is centered on the location (median) of a population.

There are two scenarios, or say, two kinds of dataset, to apply Sign Test.

- **Paired Data**: pairs of "pretreatment" and "posttreatment' observations. We would like to inference the existence of a shift in location due to the "treatment".
- **One-sample Data**: Observations from a single population about whose location we

wish to make inferences.

2.1 Paired Data

Paired Data is also known as Dichotomous Data.It consists of n independent subjects and each subject has 2 observations. The follow is an example of the paired data.

Subject i	X_i	Y_i
1	X_1	Y_1
2	X_2	Y_2
⋮	⋮	⋮
⋮	⋮	⋮
⋮	⋮	⋮
n	X_n	Y_n

Figure 1

2.2 Assumptions

Let $Z_i = Y_i - X_i, i = 1, \ldots, n$. The difference $Z_1, \ldots, Z_n$ are mutually independent, while X_i and Y_i can be dependent.

Each Z_i comes from a continuous population (not necessarily the same one) has a common median θ. The parameter θ is referred to as **the treatment effect**.

- $\mathbf{P}(Z_i \leq \theta) = \mathbf{P}(Z_i > \theta) = \frac{1}{2}$
- $\mathbf{P}(Z_i - \theta \leq \theta) = \mathbf{P}(Z_i - \theta > 0) = \frac{1}{2}$

2.3 Hypothesis Test

$$H_0 : \theta = 0$$

The null hypothesis is that each of the distributions (not necessarily the same) for the differences (post-treatment minus pre-treatment observations) has median 0, corresponding to no shift in location due to the treatment.

The **sign statistic** is the number of positive Z_i's, $i = 1, \ldots, n$.

$$T = \sum_{i=1}^{n} I_{Z_i>0}$$

,where

$$I_{Z_i>0} = \begin{cases} 1, \text{ if } Z_i > 0 \\ 0, \text{ if } Z_i \leq 0 \end{cases}$$

Under H_0, we have $\mathbf{P}(I_{Z_i>0} = 1) = \mathbf{P}(Z_i > 0) = \frac{1}{2}$. Then random variable $I_{Z_i>0}$ follows a Bernoulli distribution with $p = \frac{1}{2}$. So the sign statistic T follows a binomial$(n, \frac{1}{2})$ distribution. So $\mathbf{E}T = \frac{n}{2}$, $\mathbf{Var}T = \frac{n}{4}$.

Usually there are two kinds of test to apply:

- Exact Test: calculate p-value based on binomial distribution.
- Asymptotic Test: calculate p-value based on the central limit theorem, where the standaried sign statistic is sampled from the the stander normal distribution when the sample size is large enough ($n > 30$). $\frac{T-\frac{n}{2}}{\sqrt{\frac{n}{4}}} \to N(0,1)$ as $n \to \infty$.

2.4 Right-tailed Test

2.5 Left-tailed Test

2.6 Two-tailed Test