Statistics

Week 12: Other Statistical Methods (Chapter 14)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Information

ANOVA and HW3 solutions will be available on eDimension.

Homework assignment 4 will be available soon.

Please complete the **course survey**!

We now study *non-parametric* tests, which are tests that do not assume that the samples come from a nice distribution (e. g. normal). Bootstrapping and permutation tests are examples of non-parametric tests.

Outline

1. Sign test
2. Signed rank test
3. Runs test

Test for the median

The **sign test** is a simple test for the median. As the median is robust, it can be a better measure of centre than the mean (and is often used for income, property prices, etc).

We would like to test $H_0 : m = m_0$, where m is the true but unknown median and m_0 is a specific value.

Given data points $x_1, x_2, x_3, \ldots$, we first ignore any x_i that equals m_0. Let n be the number of data points that differ from m_0.

Next, we *count the number of* x_i's that exceed m_0, and call that number s_+. (We can also let $s_- = n - s_+$.)

Idea: reject H_0 if s_+ is 'too different' from $n/2$.

Sign test

If H_0 were true, then each $x_i > m_0$ with probability $1/2$. So S_+ follows a *binomial* distribution.

For the one-sided alternative $H_1 : m > m_0$, the p-value of s_+ is

$$\sum_{i=s_+}^{n} \binom{n}{i} 2^{-n}.$$

Reject H_0 if the p-value $< \alpha$.

The two-sided test is similar.

This is a *non-parametric* test, since it makes no assumption about the underlying distribution.

Example – hypothesis test

For the data values

$$6,\ 8,\ 9,\ 5,\ -7,\ 5,\ 3,\ -3,\ 0,\ -12,\ 3,\ 1,$$

test $H_0 : m = 0$ vs $H_1 : m > 0$, using $\alpha = 0.1$.

Answer: $s_+ = 8$, so the p-value is $\displaystyle\sum_{i=8}^{11} \binom{11}{i} 2^{-11} = 0.113$; do not reject H_0.

In *Excel*: use `binom.dist`.

Confidence interval

We can provide a crude confidence interval for the median m.

Suppose we want to find a 95% confidence interval (which turns out to be not possible, but we can get close).

We first work with S_+. Make each tail probability as close to 2.5% as possible:

$$\sum_{i=0}^{2} \binom{12}{i} 2^{-12} = \sum_{i=10}^{12} \binom{12}{i} 2^{-12} \approx 0.0193.$$

So with 96% probability, $3 \leq S_+ \leq 9$.

Next, we convert this into a 96% confidence interval for m, which is $[-3, 6)$.

Outline

Wilcoxon signed rank test

The Wilcoxon **signed rank test** takes into account the values of $d_i := x_i - m_0$, which makes it more *powerful*. However, it requires the extra assumption that the population distribution is *symmetric*.

Procedure: (1) *Rank* the d_i's in terms of absolute values (the smallest receives rank 1, etc). In case of ties, use the *average* rank.

(2) Let w_+ be the sum of the ranks of the positive d_i's.

(3) For moderate sized n, under $H_0 : m = m_0$, W_+ is approximately normal with

$$\mu = \frac{n(n+1)}{4}, \quad \sigma^2 = \frac{n(n+1)(2n+1)}{24}.$$

(n is the number of data values that differ from m_0.) Reject H_0 if w_+ is too many σ away from μ.

Proof

If H_0 were true, then the d_i which receives the kth rank is positive with probability $1/2$. Therefore

$$W_+ = \sum_{k=1}^{n} k\, X_k,$$

where X_k are iid Bernoulli random variables with parameter $1/2$. From this we can calculate its expectation and variance. Normality follows from a stronger form of the CLT.

Note: there is a test that uses a similar idea, called the Wilcoxon rank sum test (textbook Section 14.2, not in the course), which checks if two independent samples come from the same population.

Example

See spreadsheet *rank* for an example. Use *Excel*'s `rank.avg` and `if` functions.

In this example, $\mu = 60$, $\sigma^2 = 310$, $w_+ = 101$. The p-value is

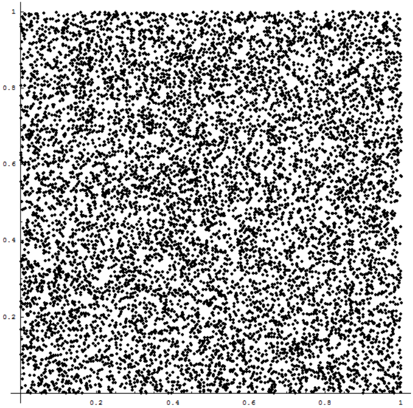
$$P(Z > \left(100.5 - 60)/\sqrt{310}\right) \approx 0.0107.$$

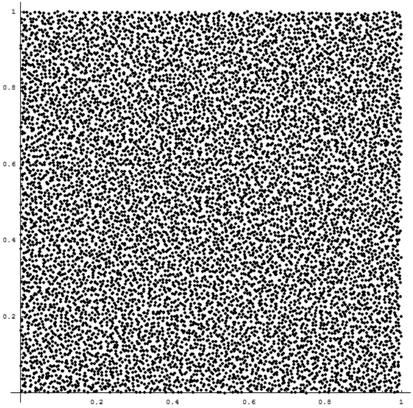
We have used the continuity correction.

Note that the *sign test* would not have rejected H_0.

Outline

Which one of the following pictures is random?





Number of runs

Suppose we observe a sequence of events, consisting of n H's and m T's (modeled as coin tosses). P(H) is *unknown*.

We would like to know if the elements of the sequence are *independent*. This could be used to check if a sequence of coin tosses were made up, or if it were random.

To do so, we count the number of **runs**.

H, T, H, H, H, H, H, T, T, H, T, T, T, H, T,
H, H, T, T, H, T, T, T, H, H, H, H, T, T, H.

T, H, H, T, H, T, H, T, H, H, T, H, H, T, H,
T, T, H, H, T, H, T, T, H, T, H, H, T, H, T.

In the first sequence, there are 8 runs of H's, in the second sequence there are 11's runs of H. (Runs of T's are *within 1* of runs of H's.)

The Wald-Wolfowitz runs test

Under H_0 : *H's and T's are independently drawn from the same distribution*, we have:

The *total* number of runs is approximately normal with

$$\mu = \frac{2mn}{m+n} + 1, \qquad \sigma^2 = \frac{(\mu-1)(\mu-2)}{m+n-1},$$

where n is the number of H's and m is the number of T's .

Thus, if the total number of runs is too many σ away from μ, then we reject H_0.

Exercise

For the second sequence of coin tosses, compute the p-value for the total runs, and determine if the sequence was made up. Use $\alpha = 0.05$, and do not forget the continuity correction.

Proof (probability for runs of H's)

We give some reasons as to why the runs are normally distributed. As a simplification, the proof here *only* considers the runs of H's.

If we assume that H's and T's are independently drawn from the same distribution, then any of the $\binom{m+n}{n}$ arrangements are equally likely. We want to know the probability of getting r runs of H's.

Consider the H's first, there are $\binom{n-1}{r-1}$ ways to break them up into r runs.

To distribute the r runs among the m T's, there are $\binom{m+1}{r}$ ways.

The runs of H's thus satisfy a *hypergeometric* distribution:

$$P(r \text{ runs of H's}) = \frac{\binom{n-1}{n-r}\binom{m+1}{r}}{\binom{m+n}{n}}.$$

Proof (normal approximation)

For moderate sized m and n, this can be approximated using a *normal* distribution with the same mean and variance:

$$\mu = \frac{n(m+1)}{m+n}, \qquad \sigma^2 = \frac{n(n-1)m(m+1)}{(m+n-1)(m+n)^2}.$$

(So for $m \approx n$, we expect around $n/2$ runs of H; anything too different allows us to reject H_0.)

Note that the Wald-Wolfowitz runs test also takes into account the runs of T's, but the proof idea is similar.

Applications

The runs test is very useful, as it makes no assumptions about P(H), and there are many ways to convert data into a sequence of H's and T's:

- Wins and losses,
- Whether data values are above/below the median,
- For integer data, whether the values are even/odd,
- Whether data values over/under-fit a distribution.

There are also tests based on the longest runs (e. g. in N tosses of a fair coin, the longest run of H's or T's is very likely to be around $\log_2(N) - 0.5$).