Before start, we recall some previous knowledges and clarify some important assumptions. Two sample T test assumptions are

- Both population involved are continuous and following normal distribution with equal variance.
- Samples are independents (if dependent there is paired t-test).
- SRS must be applied to obtain all samples.

As a non parametric method, there is no assumption, or any presumption about the parameter (i.e the shape that the distribution relays on). 比如说在参数统计推断中，我们们原假设为 $\mu_1 = \mu_2$，两个均值都有具体数值带入；而在非参数统计推断中，假设更加宽泛:

$$\mathcal{H}_0 : \text{Two population are equal.}$$

This is often interpreted as *tow population are equal in terms of their central tendency.* 只是说 distribution free 并不严谨，因为至少我们假设 samples *i.i.d*。

1. **Simplest Sign test**

   The sign test base on the test of Median. We want to test whether θ_0 is the true parameter in order to find the location of the population distribution.

   Let $\theta = \theta_0$ (where θ is a parameter and θ_0 is a particular value) and let $\{x_1, x_2, ..., x_n\}$ be the samples we get. Then with θ_0 given, we compute the series of $\theta_0 - x_i$ and record the number of $'+'$ sign indicted by t. We use t as the test statistics. Our null hypothesis is

   $$\mathcal{H}_0 : \text{ The median of X is } \theta_0$$

   and the alternative can be in three form similar to the parametric test. Also, the test statistics is whether $\theta_0 - x_i$ or the other way around depends maybe on the form of alternative.

   So under the null we can expect the test stat t follows

   $$t \sim \mathcal{B}in(n, \frac{1}{2})$$

   Or when n is large (large sample size), it if feasible to use normal distribution to approx. the binomial (since Bin. is the sum of n independent Bernoulli r.v and then by CLT) which is

   $$T = \frac{t - n/2}{\sqrt{\frac{n}{4}}} \sim \mathcal{N}(0, 1)$$

   注意，如果样本数目 n 小于 15 的话，需要做 continuous correction——即在上述 T 的分子减去 1/2。

   The sign test can be extended to a paired test analogous to the two sample t test. May make up

2. **Wilcoxon Rank sum Test (Two sample)**

   The Wilconxon rank sum test is for *two sample* and intended to for testing difference in *location* in the corresponding distributions.

   Say we have $\{x_1, ..., x_n\}$ samples from F_X and $\{y_1, ..., y_m\}$ from F_Y. Our null hypothesis is

   $$\mathcal{H}_0 : \ F_Y = F_X$$

   It is usually used for testing *slippage*, which means 概率分布相同，只是有 constant phase shift。Thus we want to test whether $\theta = 0$ in $F_Y(x) = F_X(x - \theta)$ so the hypothesis can be changed to

$$\mathcal{H}_0: \theta = 0 \ \& \ \mathcal{H}_a: \theta > / < / = 0$$

The test statistics we use is

$$W_X =: \sum_{i=1}^{n} R_i^X$$

where R_i^X is the rank of the ith sample in the sample set and W_X is simply the summation or ranks. Also we can define W_Y analogously. Notice that if finally we reject the null ($\theta = 0$), then we say X and Y has different distribution. This 'difference' is based on the location, since we assume the shape of distribution to be the same.

We have mainly 3 ways to find the distribution of test statistics.

- Enumerate the possibilities of W. 穷举出 rank 和的可能然后做比值。For example, we have $\binom{n+m}{n}$ ways of ordering for X and Y and equally likely under the null.
- Table of critical values. Not sure
- Normal approximation. If m and n are reasonably large, then under null we approximately have

$$W_X \sim \mathcal{N}(\frac{1}{2}n(n+m+1), \frac{1}{12}mn(m+n+1))$$

  proof needed, may still CLT.

If there are ties (same ranked values), use there original average rank (i.e 1,2,3,4,4, then use (4 + 5) /2 = 4.5). Notice, as in most other nonparametric test, it is not often possible to construct a test with a specific exact significance level, since the test statistics is discrete.

注意，在排序时，遇到重复数据，除了将其本身用均值代替外，不能改变之后数据的原始排序。如 $\{1,2,3,4,4,6,7\}$ 要排成 $\{1,2,3,4.5,4.5,6,7\}$。虽然对于该 test 来说没有影响因为是求和，用均值和非均值结果一样。

3. The Singed Rank Test

The singed rank test is the counterpart of matched pair t test in non-parametric test dependent?. Assume we have n math pairs$\{(x_1,y_1),...,(x_n,y_n)\}$. Under the null hypothesis, we assume the with-in pair difference is symmetric about zero (i.e $y_i - x_i$). The test statistics is constructed as follow:

- Compute the difference for each pair: $y_i - x_i$. If any zero, get rid of it and reduce n by 1.
- Take absolute vale and rank it: $|y_i - x_i|$. From lowest 1 to the highest. If any ties, assign the average rank as before.
- Sum all ranks by the rule: Let T_i be the rank of ith difference, then

$$T_+ = \begin{cases} 0, & \text{if } y_i - x_i < 0 \\ \text{rank of } |y_i - x_i|\ , & \text{if } y_i - x_i > 0 \end{cases}$$

Under the null, the difference as a random variable has 1/2 chance to be negative or to be positive, where all T_i can be thought as a r.v following the Bernoulli distribution. So if the null is true we can expect there are half positive sign in our sample. The distribution of T_+ is approximate normal, if n is large enough, with parameter

$$T_+ \sim N(\frac{n(n+1)}{4}, \frac{n(2n+1)(n+1)}{24})$$

If the sample size we have is not large enough, (maybe) the only way we can do this is by enumerating all the possibility and take the ratio (i.e # of cases that have the same result over total # of cases).

Example: 一个关于 dependent paired test 的例子。比如想要知道是否 A 燃油比 B 燃油在相同的油量下跑的里程数更多。则我们通过找到 n 辆车，让每辆车进行测试。这样得出的是 paired 的数据，因为每辆车本身会有影响，使两次测试为非独立事件。

Remark 1. The singed rank test taking into account not only the sign of difference but also their magnitude (by ranking). Also the singed rank test is good in terms of having high power.

4. **Kruskal-Wallis test**

Used to deal with samples from three or more population. Or another important application is that data (samples) are responses from a single-factor experiment. 例如把一群学生分为 n 组，每组一不同的方式教学，最后用同一个考试得到各个分数进行比较。

So we may test the viability of a null in the form

$$\mathcal{H}_0 : \text{The data in each group are from identical distribution}$$

and the alternative

$$\mathcal{H}_a : \text{The data in each group are not all from identical distribtion}$$

similar to ANOVA. The flaw of ANOVA is that each population must be normally distributed with same variance, while in non-parmatric test it is relaxed. We define the test statistics as

$$H := \frac{12\sum_{i=1}^{g} n_i(\bar{R}_i - \bar{R})^2}{n(n+1)}$$

where g is the # of groups and all ties are handled as in rank sum test. H follows a *Chi-suqare* distribution χ^2_{g-1}, since it is the sum of g standard normal distributed r.v.. The test is always one-tail. 如果统计值 H 很大，说明各个组的样本极有可能是来自不同的总体分布，因为如果是来自同样的分布总体，则排名的顺序不应该有太大的差别。H 越大则越倾向于 H_a。

5. **Permutation test**

The permutation test is very simple. The test statistics is simply the # of combinations that are exactly the same as the sample we get. The p-value can be find by the ratio of # of cases that at least as inconsistent with the null over the total # of cases. may make up later.

6. **Goodness of Fit Test**

Notice starting form here, the tests become parametric. The test statistics often used here is the *Chi-square test*. It can be different in terms of degree of fredom in *simple chi-square test* and *contingency table*.

In essence the *Chi-square test* is trying to depict the difference of the expected value and the observed value. Often the expected value comes from either the observed or theoretical null hypothesis.

In all the following test that involves using the chi-square distribution, the degree of freedom is determined as follows. The general formula is

$$df = k - 1 - p$$

We denote k as the total number of target object (i.e # of cells in a contingency table), p as the # of parameter that we need for generating the expectation of all the target object. So starting with no loss of degree of freedom which is k in total, this means we can tweak any all the value freely. Then we take into consideration the null hypothesis as a constraint. The null hypothesis usually based on either a theoretical statement (i.e the Heredity law) or a general statement as 'independent'. For the theoretical one the number of parameter is usually specific since we usually know the # parameter we need to generate the expectation (i.e in binomial, once we know n and p the whole distribution is thus determined, the # of parameter is 2 and p = 2-1 = 1), while for the latter one, we can only estimate the expectation based on over expectation: By assuming independency, the probability of being at one of the cell is the product of one of the row and column variable.

(a) **Simple Chi-square test**

In the simplest case, we have only one category (i.e performed in a table is like one row). So the test statistics is in form of

$$\sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{k-1}$$

we reject the null if the sum is too large and notice the chi-square test is alway on tail. Some notice for the test: (1) The expected value should not be too small. It should be at least 3 (2) $k-1$ is the # of free parameter, usually the # of cells - 1 (3) Sometimes we can combine several cells although it may cause the lose of power.

(b) **Contingency tables**

Most commonly use of Chi-square test. Data are cross-classified by two categorical variables (it can be classified by more). Chi-square test is used to detect whether there is any association (dependency) between categories. Assume we have a r row and c column table. The total counts is n which is known. so

$$\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij} = n$$

The our null hypothesis is

$$\mathcal{H}_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$$

then the test statistics is

$$\chi^2_{(r-1)(c-1)} := \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - np_{i\cdot}p_{\cdot j})^2}{np_{i\cdot}p_{\cdot j}}$$

由于不知道真正的分类变量的概率，所以我们用观测数据（统计量）来估计总体参数，即

$$e_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{o_{i\cdot}o_{\cdot j}}{n}$$

The degree of freedom is from

$$rc - 1 - (r - 1 + c - 1) = (r-1)(c-1)$$

(c) **Tests for Homogeneity**

(d) **Fisher's Exact test: 2 × 2 table**

Akin to the permutation test. Find out the # of ways that the table can be as extreme as the observed one. Fix all the four marginal total. The p-value is exact as the name of the text indicate. The null hypothesis is

$$\mathcal{H}_0 : \textit{The relative proportions of one variable are independent of the second variable}$$

In other words, the probability of getting r.v 1 is the same as the probability of getting r.v 2. So the test statistics under H_0 is

$$t := \frac{\binom{C_1}{o_{11}}\binom{C_2}{R_1 - o_{11}}}{\binom{n}{R_1}}$$

treat each observation as unique. It can be observed that t is actually the hypergeom. random variable. So for the *p-value*, sum up the probability of cases that are at least as possible as the observed case. We can think of the FET as a non-paramatic test, so no need to care about one or two sided.

Some remark (1) We can also use the chi-square goodness of fit test, but it can perform poorly due to the small expected counts (less than 3). So in such case the FET is more preferable. (2) For small overall sample size n, chi-square test perform poorly in terms of sig. level and power。

(e) **Normal density ploting**

Used to test whether the samples are extracting from a normal population. First we get n samples $\{x_1, x_2, ..., x_n\}$ and assume they are i.i.d r.v. with mean μ and sd σ. Then rank them from smallest to the biggest which is $\{x_{(1)}, x_{(2)}, ..., x_{(n)}\}$. We know if the samples are from normal population then it must satisfy

$$Y = \frac{X - \mu}{\sigma} \sim Normal(0, 1)$$

doing some algebra we get

$$x_{(i)} = y_{(i)}\sigma + \mu$$

which is a linear relation. make up

7. **ANOVA: a more formal approach**

Questions

1. Course note exercises 4: when use negative rank sum?
2. WW2 Q1: why for p value the alternative is "less" not grater?
3. For homoe. test: What does homogeneous population means?