Packet 7: More Testing

Chap 8.4 The Wilcoxon Test

We have learned classical inference (with parametric statistics): point estimates, confidence intervals, and hypothesis testing, regression.

We will make two extensions beyond the classic methods:

An example of Bayesian estimation: what is the probability of getting free throw points?

- Shaq O'Neal, scored two free throws, what is your belief on his FT% in this game? MLE? His historical average FT is 52.7%.
- Stephen Curry, missed two free throws, what is your belief on his FT% in this game? MLE? His historical average FT is 90.0%.

Frequentist's view: the values of parameters are fixed but unknown. After collecting more and more samples, our estimates of those parameters converge to the truth.

Bayesian's view: before a dataset is obtained, the population characteristics and the dataset are uncertain. After a dataset is obtained, the information it contains can be used to decrease our uncertainty about the population characteristics.

The basic idea of **nonparametric** statistic is to use data to infer an unknown quantity while making as few assumptions as possible.

1. Estimating the distribution function, P(X<=c);
2. Estimating the density, f(x);
3. Nonparametric regression, E(Y|X=x) = r(x);
4. Nonparametric tests.

Recall that we assume $X_i \sim N(\mu, \sigma^2)$ and test $H_0 : \mu = \mu_0$ v.s. $H_1 : \mu \neq \mu_0$.

Here, parametric means we assume that the data is generated from a known distributions, but the exact value of parameter is unknown.

For example, the normal distribution is symmetric at μ and has a bell shape. However, we may not always feel comfortable about assuming normal distributions.

E.g. Income: positive, not symmetric, (can be modeled by gamma or log-norm).

E.g. Height of adults, male height $\sim N(\mu_1, \sigma_1^2)$ and female height $\sim N(\mu_2, \sigma_2^2)$. Altogether, it has two modes, (can be modeled by mixture models).

Another possible solutions is the non-parametric hypothesis tests: distribution free tests. In non-parametric, the median is more commonly used than the mean.

The **Median** is a number such that $P(X > m) = P(X < m) = 0.5$, and m does not necessarily equal to the mean u unless for a symmetric distribution.

Sign test without distribution assumptions

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random samples obtained from a population with unknown distributions. We want to test $H_0 : m = m_0$ v.s. $H_1 : m \neq m_0$.

Note that the signs of $(X_i - m_0)$'s are also i.i.d.

Let Y be the number of negative signs among $X_i - m_0$, $i = 1, \ldots, n$. Y follows Binomial(n, p), and $p = 0.5$ if $H_0 : m = m_0$ is true.

The original hypothesis $H_0 : m = m_0$ v.s. $H_1 : m \neq m_0$ is equivalent to the hypothesis $H_0 : p = 0.5$ v.s. $H_1 : p \neq 0.5$

Recall that the test statistic for the proportion parameter is

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1),$$

and the critical region is $|Z| >= z_{\alpha/2}$.

Example 8.5-2, length of fish: $X_i = 5, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6, 1.7, 4.3$.

We want to test $H_0 : m = 3.7$ v.s. $H_1 : m \neq 3.7$.

Sign test for comparing two populations:

Suppose we have a paired sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Without assuming the distributions of X and Y, we want to test $H_0 : P(X > Y) = 0.5$ v.s. $H_1 : P(X > Y) \neq 0.5$.

For example X is the length of left foot and Y is the length of right foot of the same person.

Let Z be the number of of negative differences among $X_i - Y_i$, $i = 1, \ldots, n$,

$Z \sim$ Binomial(n, p) where $p = P(X < Y)$. It is equivalent to test

$H_0 : p = 0.5$ v.s. $H_1 : p \neq 0.5$.

Sign test for other percentiles:

We may test whether 25th percentile, $q_{0.25}$ = some fixed value, c.

Let Y be the number of observations that take values less than c.

T-test:

If the assumption $X \sim$ Norm is true, the t-test is preferable because it is more powerful than the sign test under the same significance level.

If the assumption is wrong, then t-test does not provide correct the α.

It has been criticized that a sign test does not use all the information in the observed data, for example, it does not consider the magnitude of data but only the signs.

Wilcoxon Signed Rank Test:

To consider both the sign of $X_i - m_0$ and the magnitude, $|X_i - m_0|$, but the test requires the assumption that the distribution of data is symmetric about its median.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random samples from an unknown, symmetric distribution.

1. Compute $X_i - m_0$;
2. Let R_i be the rank of $|X_i - m_0|$;
3. Determine the sign of $X_i - m_0$: 1 if $X_i - m_0 > 0$, and -1 otherwise;
4. Define $W = \sum_i sign(X_i - m_0) \times R_i$, H_0 is supported if W is close to zero.

   Here, we need to know the distribution of W so that we can control the type I error.

5. Test statistic: $Z = \frac{W}{\sqrt{n(n+1)(2n+1)/6}} \sim N(0,1)$ approximately.

Example 8.5-2: A sample of 5 fish lengths: $X_i = 5, 3.9, 5.2, 5.5, 2.8$. We want to test $H_0 : m = 3.7$ v.s. $H_1 : m < 3.7$.

Discrete Data with Ties

In the case of ties, if $X_i = m_0$ for some i, that observation is deleted, and the test is performed with a reduced sample size.

If two or more observations are equal, each observation is assigned the average of the corresponding ranks.

Example 2: Dental researchers have developed a new material for preventing cavities, a plastic sealant that is applied to the chewing surfaces of teeth. To determine whether the sealant is effective, it was applied to half of the teeth of each of 12 schoolaged children. After two years, the number of cavities in the sealantcoated teeth and in the uncoated teeth were counted, resulting in the following data:

Child	Coated	Uncoated	Diff
1	3	3	0
2	1	3	2
3	0	2	2
4	4	5	1
5	1	0	-1
6	0	1	1
7	1	5	4
8	2	0	-2
9	1	6	5
10	0	0	0
11	0	3	3
12	4	3	-1

Is there sufficient evidence to indicate that sealantcoated teeth are less prone to cavities than are untreated teeth?

Comparing two populations / distributions

Let $X_1, X_2, \ldots, X_{n_1}$ be i.i.d. samples from an unknown distribution with median m_X.

Let $Y_1, Y_2, \ldots, Y_{n_2}$ be i.i.d. samples from an unknown distribution with median m_Y.

$H_0 : m_X = m_Y$.

1. Assign ranks to the combined samples $(x_1, x_2, \ldots, x_{n_1}, y_1, y_2, \ldots, y_{n_2})$.
2. Let W be the sum of ranks from $(y_1, y_2, \ldots, y_{n_2})$.

   Under H_0, $W \sim N(\frac{n_2(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{2})$.

   Test statistics: $Z = \frac{W - n_2(n_1+n_2+1)/2}{\sqrt{n_1 n_2(n_1+n_2+1)/2}} \sim N(0,1)$.

Example 2: The weights of cinnamon packages from two companies:

X:	117.1	121.3	127.8	121.9	117.4	124.5	119.5	115.1
Y:	123.5	125.3	126.5	127.9	122.1	125.6	129.8	117.2

Examples to read 8.4-4, 8.4-5.

Mann-Whitney U statistic

$$U = W - n_2(n_2 + 1)/2$$