STAT 212: Principles of Statistics II

Lecture Notes: Chapter 4

Nonparametric Methods
('Distribution-Free' Tests)

Sharmistha Guha
Dept. of Statistics
Texas A&M University

Fall 2022

Distribution-Free Procedures

Throughout 211 and 212 we have made assumptions such as *"the data are a random sample from the Normal distribution"* and *"the errors are Normally distributed."*

The validity of the tests and confidence intervals we've discussed relies to some extent on these assumptions. *Validity definitely relies on the assumptions when the sample size is not very big.*

The motivation for *distribution-free, or non-parametric* procedures is that they retain their validity under very general conditions. In other words, *we don't have to make restrictive assumptions for these procedures to be valid.*

We'll discuss the *sign test*, the *signed-rank test* and *nonparametric ANOVA*.

Sign test

We go back to a 211 problem: testing a hypothesis about the "center" of a population.

Two common measures of center are the mean and median.

The t*-test* was used to test hypotheses about the population *mean*.

The *sign test* will be used to test hypotheses about the population *median*.

When the population is symmetric, the mean and median are, of course, the same.

If the mean and median are different, as with skewed distributions, the t-test and sign test are testing different hypotheses.

Let X be a continuous random variable with median $\tilde{\mu}$. This means that

$$P(X \geq \tilde{\mu}) = P(X \leq \tilde{\mu}) = \frac{1}{2}.$$

Suppose that $X_1, \ldots, X_n$ is a random sample from the distribution possessed by X.

Goal is to test:

$$H_0 : \tilde{\mu} = c \quad \text{vs.} \quad H_a : \tilde{\mu} > c.$$

The motivation for the sign test is as follows:

- *When H_0 is true, around 1/2 of the X_is will be larger than c.*

- *When H_a is true, the fraction of X_is larger than c will tend to be more than 1/2.*

Our test statistic will be:

$$Y = \text{number of } X_i\text{'s larger than } c.$$

We will reject H_0 when Y is 'too big'. How big is 'too big'?

Under H_0, Y has a binomial distribution with number of trials equal to n and success probability 1/2.

As long as $n \geq 10$, it is reasonable to use the Normal approximation to the binomial to carry out the test.

When H_0 is true, the distribution of

$$\frac{Y - n/2}{\sqrt{n}/2}$$

is approximately standard Normal. So, H_0 is rejected at level α if:

$$\frac{Y - n/2}{\sqrt{n}/2} \geq z_\alpha.$$

Of course, we can also test:

$$H_0 : \tilde{\mu} = c \quad \text{vs.} \quad H_a : \tilde{\mu} < c.$$

and

$$H_0 : \tilde{\mu} = c \quad \text{vs.} \quad H_a : \tilde{\mu} \neq c.$$

We use the same test statistic as before but different rejection regions. These are, respectively,

$$\frac{Y - n/2}{\sqrt{n}/2} \leq -z_\alpha$$

and

$$\frac{|Y - n/2|}{\sqrt{n}/2} \geq z_{\alpha/2}.$$

Example 20: *Distribution of pH values*

Observations: pH values of synovial fluid taken from the knees of arthritis sufferers

7.02	7.35	7.34	7.17	7.28	7.77	7.09
7.22	7.45	6.95	7.40	7.10	7.32	7.14

True median pH for nonarthritic individuals: 7.39

Does the median pH for arthritis sufferers appear to differ from that for nonarthritic individuals?

$$H_0 : \tilde{\mu} = 7.39 \quad \text{vs.} \quad H_a : \tilde{\mu} \neq 7.39$$

There are three data values larger than 7.39, so $Y = 3$. The test statistic is

$$\frac{Y - n/2}{\sqrt{n}/2} = \frac{3 - 7}{\sqrt{14}/2} = -2.13809.$$

The P-value is:

$$P = 2P(Z \geq 2.13809) = 0.0325.$$

H_0 would be rejected for any P-value ≤ 0.05. So, there is significant evidence that the median of arthritis sufferers differs from 7.39.

The sample median is 7.25, suggesting that the median pH for arthritis sufferers is less than that of nonarthritic people.

Pros and cons of the sign test

Pros

- The only assumption needed is that the data are a random sample.

- Don't need to assume anything about the distribution.

- When the population is "heavy-tailed," the sign test is usually more powerful than the t-test.

Cons

- Suppose that the population *is* Normally distributed. Then, of course, a t-test will be more powerful than the sign test.

Signed-rank test

The sign test is a nice alternative to the t-test but in a lot of situations it's not terribly powerful.

Suppose we're willing to assume that the population is *symmetric*, but not necessarily Normal.

Again we want to test:

$$H_0 : \tilde{\mu} = c \quad \text{vs.} \quad H_a : \tilde{\mu} > c.$$

Let $X_1, \ldots, X_n$ be a random sample from the population of interest.

The *signed-rank test* is a test based on the ranks of $|X_1 - c|, \ldots, |X_n - c|$.

Procedure:

(1) Rank $|X_1-c|, |X_2-c|, \ldots, |X_n-c|$ from smallest to largest, keeping track of the sign of each $X_i - c$.

(2) Define

$$I(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

and $R(|X_i - c|)$ to be the rank of $|X_i - c|$ among all of $|X_1 - c|, |X_2 - c|, \ldots, |X_n - c|$.

(3) Compute the signed-rank statistic:

$$S_+ = \sum_{i=1}^{n} I(X_i - c) R(|X_i - c|).$$

(4) Define

$$Z = \frac{S_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}.$$

When H_0 is true and $n \geq 10$, the statistic Z has approximately the standard Normal distribution. For a test with level of significance α, H_0 would be rejected if $Z \geq z_\alpha$.

Of course you could also have the alternatives

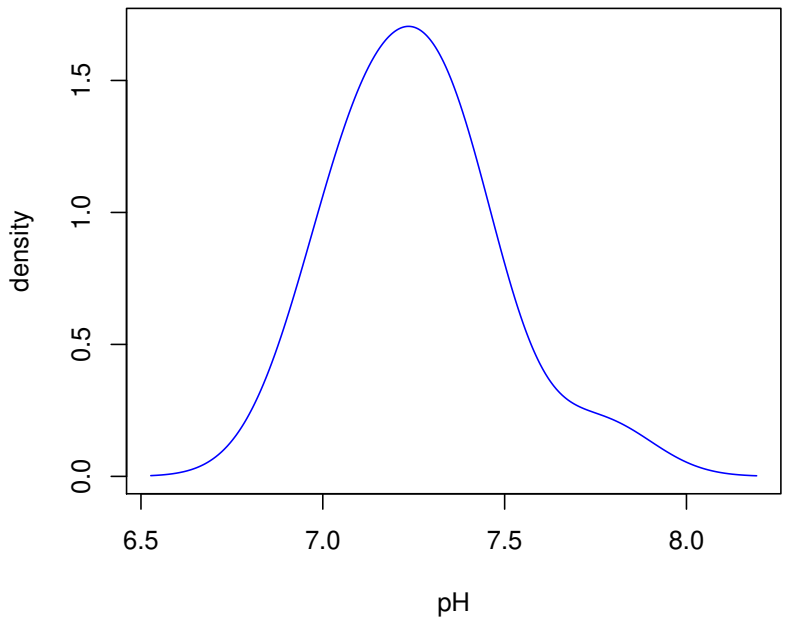
$$H_a : \tilde{\mu} < c \quad \text{or} \quad H_a : \tilde{\mu} \neq c.$$

You would still use the test statistic Z for either of these cases with respective rejection regions $Z \leq -z_\alpha$ and $|Z| \geq z_{\alpha/2}$.

- The signed-rank test uses more of the information in the data than does sign test. As a result it will often be more powerful than the sign test.

Example 21: *Signed-rank test for data in Example 20*

Kernel density estimate for the data



The kernel density estimate seems to be reasonably close to symmetric. So using the signed-rank test is ok.

Again we wish to test the hypotheses

$$H_0 : \tilde{\mu} = 7.39 \quad \text{vs.} \quad H_a : \tilde{\mu} \neq 7.39.$$

First we compute all the differences $X_i - 7.39$.

-0.37	-0.04	-0.05	-0.22	-0.11
0.38	-0.30	-0.17	0.06	-0.44
0.01	-0.29	-0.07	-0.25	

Now we rank $|X_i - 7.39|$.

12	2	3	8	6
13	11	7	4	14
1	10	5	9	

Finally, we put the appropriate signs on the ranks.

-12	-2	-3	-8	-6
13	-11	-7	4	-14
1	-10	-5	-9	

To get S_+ we add up the ranks that don't have minus signs on them.

$$S_+ = 1 + 4 + 13 = 18$$

If we take $\alpha = 0.05$, then H_0 will be rejected if

$$Z \geq 1.96 \quad \text{or} \quad Z \leq -1.96.$$

We have

$$Z = \frac{18 - 14(15)/4}{\sqrt{14(15)(29)/24}} = -2.166,$$

and so we reject H_0. *Median pH value appears to be less than 7.39 for arthritic patients.* The P-value would be $2(0.015) = 0.03$.

Distribution-free ANOVA: the Kruskal-Wallis test

Problem of interest: *Test whether several populations all have the same center.*

Assumptions:

- We have a random sample from each of k populations.

- All populations have the same distribution, except that possibly they are shifted apart.

We can have different sample sizes from the various populations. Denote these $n_1, n_2, \ldots, n_k$.

The data are

$$X_{ij}: \quad j = 1, \ldots, n_i; \ i = 1, \ldots, k.$$

Want to test the hypothesis that all k populations have the same median (or mean).

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2 = \cdots = \tilde{\mu}_k$$

$$\tilde{\mu}_i = \text{median of population } i$$

Let $N = \sum_{i=1}^{k} n_i$.

Test procedure

- Rank all N data values from smallest to largest.

- Let R_{ij} denote the rank of X_{ij} among all data.

- Define

$$\bar{R}_i = n_i^{-1} \sum_{j=1}^{n_i} R_{ij}$$

and

$$s_{KW}^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(R_{ij} - \frac{N+1}{2} \right)^2 .$$

- The test statistic is

$$KW = \frac{1}{s_{KW}^2} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2 .$$

- Reject H_0 if KW is "large."

When H_0 is true, KW has approximately the χ^2 distribution with $k-1$ degrees of freedom. So, for a level α test reject H_0 when

$$KW \geq \chi^2_{k-1,\alpha}.$$

Notes

- In order for the χ^2 approximation on the previous page to be valid, we should have each $n_i \geq 6$ when $k = 3$ or each $n_i \geq 5$ when $k > 3$.

Example 23: *Concentration of strontium-90 in milk* (the dataset can be found in Canvas)

Five milk samples are obtained from each of four regions.

The concentration of the radioactive isotope strontium-90 in each milk sample is measured.

It is of interest to compare the four regions with respect to their strontium-90 levels.

$\tilde{\mu}_i =$ median strontium-90 level for region i

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3 = \tilde{\mu}_4$$

$$\bar{R}_1 = \frac{31}{5} \quad \bar{R}_2 = \frac{68}{5} \quad \bar{R}_3 = \frac{26}{5} \quad \bar{R}_4 = \frac{85}{5}$$

$$s^2_{KW} = 35$$

$$\begin{aligned} KW &= \frac{5}{35}\Big[(6.2 - 10.5)^2 + (13.6 - 10.5)^2 + \\ &\qquad (5.2 - 10.5)^2 + (17 - 10.5)^2\Big] \\ &= 14.063 \end{aligned}$$

Since $KW = 14.063 < \chi^2_{3,0.005} = 12.838$, Reject H_0. So, there is strong evidence that the regions don't all have the same median level of strontium-90 in their milk.

It appears that regions 2 and 4 have higher levels than 1 and 3. This could be confirmed by applying the Kruskal-Wallis test to *pairs* of regions.
