From the Expert: Nonparametric Statistics

From the Expert: Nonparametric statistics

The hypothesis testing procedures discussed previously are called parametric statistics since they are based on assumptions regarding to population distribution (e.g. normal distribution), and estimation of parameters also based on samples drawn from this population. When the assumptions are not satisfied, a different technique called ***nonparametric statistics*** can be applied. Nonparametric statistics or distribution-free statistics make fewer (less strict) or no assumptions about the distribution of the data. However, the tests are less powerful since more observations are needed to obtain a conclusion with similar certainty, and they do not easily generalize to complex problems.

Nonparametric statistics is applicable when 1) the dependent variable is quantitative but its distribution is not normal or unequal population variances 2) the dependent variable is rank or ordinal 3) comparing two distributions regardless of their distributions.

Sign Test (test for a population median)

The sign test gives inference about a population median, while the t-test makes inference about a population mean. This test makes no assumption about the population distribution. It only assumes a random sample from a population.

For example, what is the typical age for Alzheimer's disease to start?

Hypothesis

(a given value)

The example shows a 2 sided test but one-sided can also be used
The idea is that if is true, the number of observed samples should be about one-half of the sample size that is above

Steps:

1. Count number of positive observations () that greater than
2. 2.1 For small sample:
   (Use the binomial distribution table with p=0.5, n= sample size)
   Compute the p-value:
   p-value = 2P[S] for two-tailed test
   Reject , if p-value is less than

3. *2.2 For large sample*:
   (use normal approximation with p=0.5, n=sample size)
   mean
   standard deviation =

Reject if or (for two tailed test)

In R, function SIGN.test( ) is in BSDA package, usage:
> SIGN.test (data)

Wilcoxon Signed Ranks Test (Compare two populations: paired difference)

It is suitable when data measured as a pair of subjects (e.g. before and after score). In addition, the sign (positive, negative) and magnitude of the difference between matched pairs are known. It is used as a nonparametric alternative to a repeated measures t-test. Although this test does not assume normality, it requires **symmetric** distribution. Symmetric means left side of the distribution is the reflection of the right side.
Hypothesis:
: the mean difference of the pairs is equal zero)
: (the mean difference of the pairs is not equal zero)

Steps:

1. Compute the score difference for each pair and take only absolute value

2. Rank the absolute value from step1 (if 0: exclude, if ties: give an average rank)

3. Affix the sign of the score back to the rank

4. Calculate the sum ranks both the positive ranks (), and negative rank (). is the smaller of or
   4.1 For large sample size :
   Calculate the normal approximation using the following formula
   is the number of pairs
   and 24 are constants

   Using the normal table to find the probability, and then compare with (e.g.0.05)
   Rejection region:

   4.2 For small sample size:
   Calculate the sum ranks both the positive ranks (), and negative rank ().
   Test statistic: , the smaller of or .
   Reject region:
   Where is given in the table with n (# of pairs) and given
   Rejecting the null hypothesis suggests that a real difference exists between pairs.

For the wilcoxon signed rank test in R, use function wilcox.test ( ).
> wilcox.test (data, mu =0, conf.int=TRUE) # with continuity correction

Mann-Whitney U-Test (compare distribution of two independent samples)

This test is also known as the Wilcoxon rank sum test. This is the inference for two independent samples test. The required assumptions include populations being compared assumed to have the same shapes and spreads, the two samples are independent, observations in each sample must be independent of one another.

For example, we want to compare the sale performance of Team A and Team B. The revenue observations are collected from each term. Is there any difference between the sale performance of Team A and Team B?

Hypothesis

: (There is no difference in the sale performance of Team A and Team B)
: (There is difference in the sale performance of Team A and Team B)

Test statistic (:

(mean of) =

(standard deviation of)
is the number of values from the first sample
is the number of values from the second sample

When both sample sizes are more than 10, approximate t.
Steps:

1. Combine two samples of size into one sample of size
2. Sort the result, assign ranks from smallest to largest (e.g. smallest = 1) to the sorted values (if tied gives the average rank)
3. From step 2, sum rank of the first sample , and sum rank of the second sample

Calculate

Then take the smaller between and and named it
(if these two samples have the same medians, sum ranks of both samples should be very close)

1. Calculate the test statistic using the above formula and from step 3
2. Rejection region:

This suggests a difference between Team A and Team B

In R, usage:
```
> wilcox.test(x~y, data)
```

Kruskal-Wallis Test (compare three or more populations)

This is a non-parametric test for three or more independent samples. Required conditions are *k* samples that are random and independent, the samples come from continuous probability distribution, and each sample should have at least 5 measurements.

The *k* probability distributions are identical
At least two of *k* probability distributions are different

For example, we want to compare the average starting salary for 4 different majors: Sciences, Social Sciences, Education, and Humanities.
The average starting salaries for different majors are the same
The average starting salaries for different majors are different

Test statistic:

Reject , if with (k-1) degrees of freedom

Steps:
Let be the sample sizes of *k* independent samples.

1. Combine all samples (into one large sample of size

2. Sort the result from smallest to largest

3. Assign rank to the sorted result. If ties, assign the average rank

4. Sum ranks () for each group *i*th sample

5. Calculate from the above formula

6. Reject , if (e.g. the average starting salaries for different majors are different)

Kruskal-Wallis in R, usage:
> kruskal.test( )

Kolmogorov-Smirnov (K-S) Test

The K-S test can be used for a single sample (e.g. goodness of fit test) and two independent samples test (e.g. test two samples whether they come from the same distribution).

The K-S test has been widely used in many (previous) research literatures. The strength of the test include distribution free, no restriction to any scientific problems or size of the sample, and one-sample K-S test can serve as a goodness of fit test. However, the K-S test also has some limitations such as the underlying distribution is assumed to be continuous, cumulative probability is more sensitive near the center than the tails, and distribution parameters should completely specified, otherwise, the critical region of K-S test is no valid.

An example of one sample, for example, if we want to test the hypothesis that data (unknown distribution) is equal to a particular distribution.

The data follow a specified distribution
The data does not follow the specified distribution

Steps: (one sample)

1. Calculate the cumulative frequency of the observation data (add up the observed value and divide by number of sample size)

2. Calculate the cumulative frequency of the hypothesized distribution (called expected cumulative frequency)

3. Differ the observed cumulative and expected cumulative frequency.

4. Find the largest absolute difference from step 3)

5. Compare value from 4) with the critical value

6. If the largest absolute difference is greater than the critical value, we reject the null hypothesis (e.g. data does not fit the specified distribution)

An example of two independent samples: does the data support that the salary for males is the same as for females?

Steps (two independent samples)

Steps (two independent samples)
In the same manner as steps for one sample

1. Calculate the cumulative frequency of the observed data sample 1, and observed data sample 2 (add up the observed value and divide by number of sample size)
2. Differ the observed cumulative frequency of data sample 1 and data sample 2
3. Find the largest absolute difference from step 2)
4. Compare the value from 3) with the critical value
5. If the largest absolute difference is greater than the critical value, we reject the null hypothesis (e.g. there is difference in these two distributions)

K-S is considered has less power (detect a difference when it actually exists) than Shapiro-Wilk test, or Anderson-Darling test.

K-S test in R, usage:
> ks.test (x,y,.., alternative = c("two.sided","less","greater"), exact=NULL)

Anderson-Darling Test

This test measures how well the data follow a specific distribution. This is the extension of K-S test where more weights given to the tails than the K-S test. In addition, Anderson-Darling calculates critical values for each distribution, contrary to the K-S test where the critical value does not rely on the being tested distribution. Therefore, Anderson Dalring is a more sensitive test than the K-S test. At present, critical values tables are available for normal, uniform, lognormal, exponential, Weibull, etc

Hypothesis:
The data follow a specified distribution
The data does not follow the specified distribution

Anderson-Darling test in R is in package 'nortest', usage:
> ad.test(x)

Specific normality tests are the Shapiro-Wilk Test (Shapiro.test), cramer-von Mises (cvm.test), etc.

Check more for details: http://cran.r-project.org/web/packages/nortest/nortest.pdf

Spearman's Rank Correlation Coefficient

This test measures the correlation between two ranked variables. To be valid, this test requires that two variables are randomly selected and the probability distribution of both variables are continuous.

For example, we want to test whether there is any relationship between the numbers of hours spent on TV and test scores?
(There is no association between rank pairs)
(There is an association between rank pairs)

Note that the alternative hypothesis can be a one-tailed or two-tailed test. The above example is two-tailed hypothesis.

Test statistics:

Reject if , when is the value from the table

Steps:

1. Rank data in x (e.g. number of hours) from 1 to n (n is sample size), and rank y data (e.g. test score)

1. Rank data in x (e.g. number of hours) from 1 to n (n is sample size), and rank y data (e.g. test score)

2. In the same data column, subtract the value and then square the value ( )

3. Sum the result (squared) values from step 2)

4. Reject if , when is the value from the table

Several alternatives for using spearman
> cor.test (,method=”spearman”)
> spearman.test ( ) # in package ‘pspearman’

R for non-parametric statistics examples:

http://www.stat.fsu.edu/~jfrade/HOMEWORKS/STA5507/test2/R_codes_for_chapter_10.pdf
http://www.r-tutor.com/elementary-statistics/non-parametric-methods
http://www.statmethods.net/stats/nonparametric.html

References:

Boslaugh, S. (2013). Statistics in a Nutshells. 2nd Edition. O’Reilly Media. Inc..

Lim, KH. (2006) Statistics for Psychology and Human Science: Oxford University Retrieved from:
http://www.stats.ox.ac.uk/~lim/Teaching_files/Lecture5/Lecture5%20-%20Non-Parametric%20Tests%201.pdf
http://www.stats.ox.ac.uk/~lim/Teaching_files/Lecture6/Lecture6%20-%20Non-Parametric%20Tests%202(4up).pdf

Miller, R. (2014) Department of Statistics, Florida International University. Retrieved from: http://www2.fiu.edu/~millerr/Chapter%20Fourteen.pdf

Reflect in ePortfolio Download Print

Activity Details

Task: View this topic