R Short Course | MBBC Students 18/19

Statistics 1

Non-Parametric Inference

Lisete Sousa
lmsousa@fc.ul.pt
Room: 6.4.25

Departamento de Estatística e Investigação Operacional
Centro de Estatística e Aplicações

Ciências ULisboa

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

2019

Summary

1. Tests for Measures of Central Location

This section concerns one type of non-parametric procedure, tests that can replace the t-tests and one-way analysis of variance when the Normality conditions for those tests are not met.

The most useful non-parametric tests are based on the rank (place in order) of each observation in the set of all the data.

The rank tests we will study concern the **center** of a population or populations. When a population has at least roughly a Normal distribution, we describe its center by the mean. When distributions are strongly skewed, we often prefer the median to the mean as a measure of center. In simplest form, the hypotheses for rank tests just replace mean by median.

- Non-parametric alternative to `t.test` and `z.test`: `wilcox.test` (Wilcoxon test and Wilcoxon-Mann-Whitney test).
- Non-parametric alternative to `aov`/`anova`: `kruskal.test` (Kruskal-Wallis test).
- Non-parametric alternative to `var.test`: `mood.test` and `ansari.test`.
- For testing k-sample homogeneity of variances: `bartlett.test` (sensitive to departures from Normality) and `fligner.test` (non-parametric).

Wilcoxon Test

One sample / Matched pairs

We use the one-sample procedure for inference about the mean of one population or for inference about the mean difference in a matched pairs setting. We will now meet the Wilcoxon test, or Wilcoxon signed-rank test for matched pairs and single samples. The matched pairs setting is more important because good studies are generally comparative.

Example 1: Vitamin loss in a food product

Food products are often enriched with vitamins and other supplements. Does the level of a supplement decline over time, so that the user receives less than the manufacturer intended?

Here are data on the vitamin C levels (milligrams per 100 grams) in wheat soy blend, a flour-like product supplied by international aid programs mainly for feeding children. The same 9 bags of blend were measured at the factory and five months later in Haiti.

Bag	1	2	3	4	5	6	7	8	9
Factory	45	32	47	40	38	41	37	52	37
Haiti	38	40	35	38	34	35	38	38	40

We suspect that vitamin C levels are generally higher at the factory than they are five months later.

We would like to test the hypotheses:

H_0: vitamin C has the same distribution at both times

H_1: vitamin C is systematically higher at the factory

Because these are matched pairs data, we base our inference on the **differences**.

Positive differences indicate that the vitamin C level of a bag was higher at the factory than in Haiti.

If factory values are generally higher, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction (one-sided test).

We therefore compare the **absolute values** of the differences, that is, their magnitudes without a sign.

Bag	1	2	3	4	5	6	7	8	9
Factory	45	32	47	40	38	41	37	52	37
Haiti	38	40	35	38	34	35	38	38	40
Difference	7	-8	12	2	4	6	-1	14	-3
Absolute value	7	8	12	2	4	6	1	14	3
Rank	6	7^-	8	2	4	5	1^-	9	3^-

Notes:

1. Tied values receive the average of their ranks.
2. If there are zero differences, discard them before ranking.

```
> x<-c(45, 32, 47, 40, 38, 41, 37, 52, 37)
> y<-c(38, 40, 35, 38, 34, 35, 38, 38, 40)
> wilcox.test(x, y, alternative = "greater", paired=TRUE)

        Wilcoxon signed rank test

data:  x and y
V = 34, p-value = 0.1016
alternative hypothesis: true location shift is greater than 0
```

In this example p-value=0.1016, thus H_0 is not rejected for the usual significance levels. This small sample does not give convincing evidence of vitamin loss.

Mann-Whitney-Wilcoxon

Two samples

Although we emphasize the matched pairs setting, W^+ can also be applied to a single sample. It then tests the hypothesis that the population median is zero.

To test the hypothesis that the population median has a specific value m, apply the test to the differences $X_i - m$. For matched pairs, we are testing that the median of the differences is zero.

The Mann-Whitney-Wilcoxon, or Wilcoxon rank-sum test, is a non-parametric alternative to the one sample t-test which is based solely on the order in which the observations from the two samples fall. We will use the following as a running example.

Example 2: Genetic inheritance

In a genetic inheritance study discussed by Margolin (1988), samples of individuals from several ethnic groups were taken. Blood samples were collected from each individual and several variables measured.

We shall compare the groups labelled "Native American" and "Caucasian" with respect to the variable MSCE (mean sister chromatid exchange). The data is as follows:

Native American (A)	8.50	9.48	8.65	8.16	8.83	7.76	8.63		
Caucasian (C)	8.27	8.20	8.25	8.14	9.00	8.10	7.20	8.32	7.70

Here, we want to test if MSCE distribution for Native Americans is the same as that for Caucasians. Although the Native American MSCE values in the data tend to be higher, there was no prior theory to lead us to expect this so we should be doing a two-sided test.

Hypotheses:

H_0: MSCE has the same distribution for both ethnic groups

H_1: MSCE has different distribution for both ethnic groups

The Wilcoxon test is based upon ranking the $n_A + n_C$ observations of the combined sample.

Tied values receive the average of their ranks.

Native American (A)	Caucasian (C)	Rank
	7.20	1
	7.70	2
7.76		3^A
	8.10	4
	8.14	5
8.16		6^A
	8.20	7
	8.25	8
	8.27	9
	8.32	10
8.50		11^A
8.63		12^A
8.65		13^A
8.83		14^A
	9.00	15
9.48		16^A

```
> x<-c(8.50, 9.48, 8.65, 8.16, 8.83, 7.76, 8.63)
> y<-c(8.27, 8.20, 8.25, 8.14, 9.00, 8.10, 7.20, 8.32, 7.70)
> wilcox.test(x, y, alternative = "two.sided", paired = FALSE)

        Wilcoxon rank sum test
data:  x and y
W = 47, p-value = 0.1142
alternative hypothesis: true location shift is not equal to 0
```

Kruskal-Wallis test

The Kruskal-Wallis test is a rank test that can replace the ANOVA F test. The assumption about data production (independent random samples from each population) remains important, but we can relax the Normality assumption. We assume only that the response has a continuous distribution in each population.

Data considered in the Example 7, of the previous section, will be used to illustrate this test.

Example 3: Cows' weights

Four samples of weights (kg) of cows, according to the treatments A, B, C and D.

Hypotheses:

H_0: Weights have the same distribution in all groups

H_1: Weights are systematically higher in some groups than in others

The null hypothesis is that all four samples have the same median weights. The alternative hypothesis is that not all four median weights are equal.

Like the Wilcoxon rank-sum statistic, the Kruskal-Wallis statistic is based on the sums of the ranks for the groups we are comparing. The more different these sums are, the stronger is the evidence that responses are systematically larger in some groups than in others.

Treatment					Treatment				
A	B	C	D	Rank	A	B	C	D	Rank
	21.78			1				30.40	16
	25.59			2				30.70	17
		25.85		3				30.83	18
		26.22		4		32.55			19
	26.86			5			33.40		20
		27.38		6				33.54	21
27.58				7				33.84	22
27.91				8		34.26			23
			28.82	9		35.08			24
		29.11		10			39.47		25
29.33				11					
			29.60	12					
		30.15		13					
30.28				14					
		30.39		15					

```
> a<-c(30.28, 27.5, 27.9, 29.33)
> b<-c(34.26, 32.55, 21.78, 25.59, 35.08, 26.86)
> c<-c(39.47, 30.15, 33.40, 27.38, 30.39, 25.85, 29.11, 26.22)
> d<-c(33.54, 30.40, 29.60, 28.82, 30.70, 30.83, 33.84)
> x<-list(a,b,c,d)
> kruskal.test(x)

        Kruskal-Wallis rank sum test

data:  x
Kruskal-Wallis chi-squared = 2.3807, df = 3, p-value = 0.4972
```

2. The Chi-Squared Test

The Chi-Squared Goodness of Fit Test

A hypothesis test concerning the adequacy of a model is called goodness of fit test.

With this example we will see how to use the Pearson χ^2 statistic (a statistic which has an approximate χ^2 distribution) to test the adequacy of a model.

The null hypothesis here is that the variable of interest, say X, has a distribution $F_X(x)$. This distribution may be completely specified, or may be specified up to some unknown parameters. If this is the case then the parameters have to be estimated by the maximum likelihood method using the available data.

To apply the χ^2 goodness of fit test we follow the procedure:

1. Divide the support of the hypothesized distribution in k mutually exclusive and exhaustive intervals, $I_1, \ldots, I_k$.
2. From a random sample of size n, denote by O_i, $(i = 1, \ldots, k)$ the frequency of occurrence of data points in each interval.
3. Compute the expected number of points in each interval according to the hypothesized distribution, ie, $E_i = nP(X \in I_i) = np_i$.
4. Compute the statistic

$$\mathbf{x}^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

5 Under the null hypothesis and if, at least 80% of the $E_i \geq 5$, then the statistic $\mathbf{X}^2$ is approximately distributed as a χ^2 with the number of degrees of freedom equal to $k - 1 - r$, where r is the number of parameters estimated from the data.

6 Reject the null hypothesis at significance level α if the observed value for the $\mathbf{X}^2$ statistic is bigger than $\chi^2_{k-1-r}(1-\alpha)$, the quantile of probability $1-\alpha$ of a χ^2 distribution with $k-1-r$ degrees of freedom.

Example 4: Long Repeats

Suppose that we observed a very long DNA sequence and counted the number of A's before another nucleotide appears obtaining the following data:

0	2	0	0	0	1	0	3	0	1	0
0	0	0	2	0	0	0	0	0	1	0
0	0	1	0	0	3	0	1	2	0	0
0	0	0	0	0	0	0	0	1	0	2
0	0	1	0	1	0	0	1	1	0	1
0	0	1	0	0	0	0	1	0	0	1
1	1	0	0	1	0	2	0	1	0	3
2	0	0	0	1	1	0	2	0	0	1
1	0	0	0	1	3	3	0	0	0	0
1	1	0	0	1	1	4	0	0	0	1
1	0	0	5	0	1	0	0	2	0	0
1	0	0	0	0	1					

Are these data compatible with the hypothesis that the distribution of X, the length of consecutive A's, is geometric, i.e.,

$$P(X = x|\theta) = \theta(1-\theta)^x, \quad x = 0, 1, 2, \ldots$$

where $1 - \theta$ is the probability of an A in any position of the DNA sequence?

The support of the geometric distribution is $S = \{0, 1, 2, \dots\}$.

Given the data we may construct the following table of observed frequencies and expected frequencies according to the hypothesized geometric model

x_i	observed frequency (O_i)	expected frequency (E_i)	$(O_i - E_i)^2 / E_i$
0	80	81.28	0.020
1	32	29.26	0.256
2	8	10.53	0.610
≥ 3	7	5.93	0.195

To compute the expected frequencies is necessary to estimate θ from the data.

The maximum likelihood estimate of θ is $\hat{\theta} = n/(n + \sum x_i) = 0.64$. Hence the probability of an A is estimated as 0.36.

The value of the $\mathbf{X}^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$ statistic is 1.081.

The obtained p-value is 0.582, meaning that there is strong evidence that the model is adequate. In fact, we can never say that the model is "the true model". The fact that we do not reject a specific model does not even mean that it is the best model. It is only a possible model.

Note: For the calculation of the p-value we consider a χ^2 distribution with $4 - 1 - 1 = 2$ degrees of freedom since one parameter has been estimated.

```
> x<-c(0,1,2)
> obs.freq<-c(80,32,8,7)
> prob<-dgeom(x,0.64)
> prob[4]<-1-sum(prob)
> exp.freq<-sum(obs.freq)*prob
> exp.freq

[1] 81.280000 29.260800 10.533888  5.925312

> res<-chisq.test(obs.freq,p=prob)
> res

        Chi-squared test for given probabilities
data:  obs.freq
X-squared = 1.081, df = 3, p-value = 0.7817

> 1-pchisq(res$statistic,4-1-1)

X-squared
0.5824514
```

The Chi-Squared Test of Homogeneity

Suppose that we have observed the frequencies of the four nucleotides A, G, C, T in two DNA sequences

	A	G	C	T	Total
sequence 1	273	258	233	236	1000
sequence 2	281	244	246	229	1000
Total	554	502	479	465	2000

The question of interest here is the following:

"Is it true that the two sequences are drawn from 'populations' with identical nucleotide frequencies?"

To simplify notation let us codify the nucleotides in the following manner: A, B, C, B are codified as $1, 2, 3, 4$, respectively.

If we denote by p_{ij} the probability of occurrence of nucleotide j in the i sequence ($i = 1, 2$), the previous question can be translated into a hypothesis testing framework as

$$H_0 : p_{1j} = p_{2j}, \forall j = 1, 2, 3, 4 \quad v.s. \quad H_1 : \exists j \quad p_{1j} \neq p_{2,j}$$

This is a test of homogeneity of populations.

Example 5:

Representing by X_{ij} the random variable which counts the number of nucleotides of type j in the i sequence, and by E_{ij} the expected number of nucleotides of type j in the i sequence (under the null hypothesis), we can the construct the $\mathbf{X}^2$ as before

$$\mathbf{x}^2 = \sum_{i=1}^{2}\sum_{j=1}^{4}\left[\frac{(X_{ij} - E_{ij})^2}{E_{ij}}\right]$$

and compare the observed value with the adequate quantile of a χ^2 distribution with we have $(r-1)\times(c-1) = 1\times 3 = 3$ degrees of freedom.

In the present case we have the table with the observed frequencies and, in brackets, the expected frequencies under the null hypothesis:

	x_{i1}	x_{i2}	x_{i3}	x_{i4}	Total
sequence 1	273 (277)	258 (251)	233 (239.5)	236 (232.5)	1000
sequence 2	281 (277)	244 (251)	246 (239.5)	229 (232.5)	1000
Total	554	502	479	465	2000

The observed value of $\mathbf{X}^2$ is 0.964 and the p-value is 0.8099, showing a strong evidence not to reject the hypothesis that the two sequences are homogeneous.

```
> table<-matrix(c(273,281,258,244,233,246,236,229),2,4)
> dimnames(table)<-list(c("seq1","seq2"),c("A","G","C","T"))
> chisq.test(table)

        Pearson's Chi-squared test
data:  table
X-squared = 0.9642, df = 3, p-value = 0.8099

> chisq.test(table)$expected

        A   G     C     T
seq1 277 251 239.5 232.5
seq2 277 251 239.5 232.5
```

The Chi-Squared Test of Independence

.

The χ^2 test statistic is commonly used to test the null hypothesis that two criteria (A and B) of classification, when applied to the same set of entities, are independent. Data are organized in the form of two-way table, with r rows and c as it was previously seen.

If we represent by p_{ij} the probability that an entity falls in i category of B and j category of A, $p_{i\bullet}$ the probability that an entity falls in i category of B and $p_{\bullet j}$, the probability that an entity falls in j category of A, then the hypothesis of independence is

$$H_0 : p_{ij} = p_{i\bullet} p_{\bullet j}, \quad \forall i, j$$

The alternative is that there exists at least one pair (i, j) for which the joint probability is different from the product of the marginals.

The test statistic is again

$$\mathbf{X}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(X_{ij} - E_{ij})^2}{E_{ij}}\right],$$

which, under the null hypothesis of independence, follows approximately a χ^2 distribution with $(r-1)\times(c-1)$ degrees of freedom.

Example 6: A statistical test for Markov independence

Nucleotides at adjoining DNA sites are often dependent, and a first-order Markov model fits data significantly better than the independence model. A statistical test for Markov independence is an association test in a 4×4 contingency table.

Suppose that X_{ij} is the number of times, in a DNA sequence of interest, that a nucleotide of type j is followed by a nucleotide of type i in a sequence of length N. Then we can use the χ^2 test described before to test if there is association of nucleotides or not.

What conclusion would you have drawn if you had observed the following data?

		nucleotide at site k+1				
		A	G	C	T	Total
nucleotide at site k	A	67	70	55	48	240
	G	80	60	70	65	275
	C	58	72	80	66	276
	T	60	74	60	82	276
	Total	265	276	265	261	1067

The table of expected frequencies under the hypothesis of independence is:

	A	G	C	T	Total
A	59.61	62.08	59.61	58.70	240
G	68.30	71.13	68.30	67.27	275
C	68.55	71.39	68.55	67.51	276
T	68.55	71.39	68.55	67.51	276
Total	265	276	265	261	1067

The observed value of $\mathbf{X}^2$ is 17.01 and p-value 0.04851. We would reject the hypothesis of independence at the 5% level but not at the 1% level.

```
> table<-matrix(c(67,80,58,60,70,60,72,74,55,70,80,60,48,65,66,82),4
> dimnames(table)<-list(c("A","G","C","T"),c("A","G","C","T"))
> chisq.test(table)

        Pearson's Chi-squared test

data:  table
X-squared = 17.0132, df = 9, p-value = 0.04851

> chisq.test(table)$expected

          A         G         C         T
A 59.60637 62.08060 59.60637 58.70665
G 68.29897 71.13402 68.29897 67.26804
C 68.54733 71.39269 68.54733 67.51265
T 68.54733 71.39269 68.54733 67.51265
```