BioSys PhD | Earthsystems PhD

Statistics 1

Non-Parametric Inference

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

Departamento de Estatística e Investigação Operacional
Faculdade de Ciências da Universidade de Lisboa

2017

Summary

1. Tests for Measures of Central Location
   - Introduction
   - Wilcoxon Test
   - Mann-Whitney-Wilcoxon
   - Kruskal-Wallis test
2. The Chi-Squared Test
   - Goodness of Fit
   - Homogeneity
   - Independence
   - Parametric vs Non-Parametric Tests
3. The Multiple Testing Problem
   - Types of error control
   - FWER
   - FDR
   - Some final considerations

1. Tests for Measures of Central Location

Introduction

The most commonly used methods for inference about the means of quantitative response variables assume that the variables in question have Normal distributions in the population.

Fortunately, our usual methods for inference about population means (the one-sample and two-sample procedures and analysis of variance) are quite robust. That is, the results of inference are not very sensitive to moderate lack of Normality, especially when the samples are reasonably large.

What can we do if plots of the data suggest that the population is clearly not Normal, especially when we have only a few observations? This is not a simple question.

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.

(1) 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.

(2) Test statistic:

$$W^+ = \sum \textit{positive ranks}$$

The test statistic is the sum of the ranks of the positive differences. This is the *Wilcoxon signed rank statistic*. Its value here is $w^+ = 34$. (We could equally well use the sum of the ranks of the negative differences, which is 11.)

(3) Decision:

Reject H_0 for large values of the statistic W^+, since the alternative hypothesis is "vitamin C is systematically higher at the factory".

H_0 is rejected for a significance level α, if p-value$= P(W^+ \geq 34) < \alpha$. P-values can be found from special tables or software like R.

```
> 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.

(1) 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

(2) Test statistic:

The Wilcoxon rank-sum test statistic is the sum of the ranks for observations from one of the samples. Let us use w_A for the observed rank sum and W_A to represent the corresponding random variable:

$$W_A = \sum \textit{ ranks for observations from } A$$

The sum of the ranks for the Native American group is

$$w_A = 3 + 6 + 11 + 12 + 13 + 14 + 16 = 75$$

(3) Decision:

For the two-sided test, i.e., testing H_0: $median_A = median_C$ versus the alternative H_1: $median_A \neq median_C$, a rank sum that is either too big or too small provides evidence against H_0.

The rank sum for for the Native American group was $w_A = 75$. We know from the ranks table that this will be in the upper tail of the distribution. The $p - value$ is thus

$$p - value = 2 \times P(W_A \geq 75) = 0.114$$

For the usual significance levels we do not reject H_0. This suggests that median MSCE measurements are similar for Native Americans than for Caucasians.

```
> 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
```

Note:

R computes the value of the statistic as the number of all pairs (x[i], y[j]) for which y[j] is not greater than x[i]. Then, R gives as output

$W = 2 + 4 + 4 \times 8 + 9 = 47$ since

- nr of elements of C not greater than $7.76 = 2$
- nr of elements of C not greater than $8.16 = 4$
- nr of elements of C not greater than $8.50 = 8$
- nr of elements of C not greater than $8.63 = 8$
- nr of elements of C not greater than $8.65 = 8$
- nr of elements of C not greater than $8.85 = 8$
- nr of elements of C not greater than $9.48 = 9$

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.

(1) 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					

(2) Test statistic:

There are N observations in all:

$$N = n_A + n_B + n_C + n_D = 4 + 6 + 8 + 7 = 25.$$

After ranking all N observations, let R_i be the sum of the ranks for the ith sample: $R_A^2 = 1600$; $R_B^2 = 5476$; $R_C^2 = 9216$; $R_D^2 = 13225$.

The Kruskal-Wallis statistic is:

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

The observed value of the Kruskal-Wallis statistic for this example is $H = 2.38$.

(3) Decision:

The exact distribution of the Kruskal-Wallis statistic H under the null hypothesis depends on all the sample sizes, so tables are awkward. The calculation of the exact distribution is so time-consuming for all but the smallest problems that even most statistical software uses the chi-squared approximation to obtain p-values.

Here, the p-value is 0.4972, according to a chi-squared distribution with 3 (4-1) degrees of freedom.

For the usual significance levels we do not reject H_0. This suggests that there are no differences between treatments.

```
> 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

In the Example 1 of Section 3, we might have had doubts whether the geometric distribution was appropriate to model the length of sequences of consecutive A nucleotides.

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
```

Parametric tests

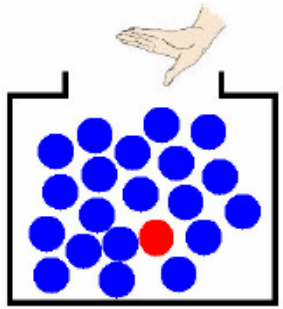
- Only for quantitative data.
- Assume that the data come from a population which follows a certain distribution (Normal distribution).
- Assuming equal variances and unequal variances.
- More powerful.

Non-parametric tests:

- Does not assume Normal distribution
- No variance assumption
- Decrease effects of outliers (robust)
- Not recommended if there is less than 5 replicates per group
- Less powerful

3. The Multiple Testing Problem

The multiple comparison problem



- Imagine a solution with 20 spheres: 19 are blue and 1 is red. What are de odds of randomly sampling the red sphere by chance? It is 1 out of 20.
- Now let's say that you get to sample a single sphere (and put it back into the solution) 20 times. Have a much higher chance to sample the red sphere. This is exactly what happens when testing several thousand tests at the same time.

- V: r.v. which represents the number of false positives on n tested hypotheses,
  $P(V \geq 1) = 1 - P(V = 0) = 1 - (1 - \alpha)^n$,
  where α is the probability of rejecting the null hypothesis when it is true (*Type I Error*), $P(\text{Rej } H_0 | H_0 \text{ True})$.

Number of hypotheses tested (n)	False positives incidence ($n \times \alpha$)	Probability of 1 or more false positives by chance ($1 - (1 - 0.05)^n$)
1	$1/20 = 0.05$	0.050
2	$2 \times (1/20) = 0.1$	0.098
20	$20 \times (1/20) = 1$	0.642
100	$100 \times (1/20) = 5$	0.994

Problem: When many hypotheses are tested, the probability of a type I error (false positive) increases sharply with the number of hypotheses.

The multiple comparison problem

in gene expression data

- There is a serious consequence of performing statistical tests on many genes in parallel, which is known as multiple testing problem.
- How do we know that the genes that appear to be differentially expressed are truly differentially expressed and are not just artefact introduced because we are analyzing a large number of genes?
- Is this gene truly differentially expressed, or could it be a false positive result?
- In such studies, the probability of at least one false positive result is near certain.

- We will often be interested not just in the probability of one error, but in the expected total number of errors. The expected number of false positives is simply α multiplied by the number of tests.

- **E-value (expected value)**: For m=100 independent tests with $\alpha = 0.1$, the expected number of false positives is $100 \times 0.10 = 10$ false positives.

Types of error control

	Rejected H_0	Not Rejected H_0	
True H_0	V	U	m_0
False H_0	S	T	m_1
	R	$m - R$	m

V: false positives (type I error)

T: false negatives (type II error)

- Suppose one test of interest has been conducted for each of m genes in a microarray experiment.
- Let $p_1, p_2, \ldots, p_m$ denote the p-values corresponding to the m tests.

- Let $H_{01}, H_{02}, ..., H_{0m}$ denote the null hypotheses corresponding to the m tests.
- Suppose m_0 null hypotheses are true and m_1 null hypotheses are false.
- Let c denote a value between 0 and 1 that will serve as a cutoff for significance:
  - Reject H_{0i} if $p_i \leq c$ (declare significant difference in the expression)
  - Do not reject H_{0i} if $p_i > c$ (declare non-significant difference in the expression)

- **PCER:** Per-comparison error rate, the expected value of the number of Type I errors over the number of hypotheses, PCER=E(V)/m.
- **PFER:** Per-family error rate, the expected number of Type I errors, PFER=E(V).
- **FWER:** Family-wise error rate: the probability of at least one type I error, FWER=P(V$\geq$ 1).
- **FDR:** False discovery rate, is the expected proportion of incorrectly rejected null hypotheses, FDR=E(V/R) for R>0.

FWER - Family Wise Error Rate

1. Many procedures have been developed to control the Family Wise Error Rate (the probability of at least one type I error): $P(V \geq 1)$
2. Two general types of FWER corrections:
   1. Single Step: equivalent adjustments made to each p-value.
   2. Sequential: adaptive adjustment made to each p-value.

Single-step approach: Bonferroni

- The Bonferroni Method is the simplest way to achieve control of the FWER at any desired level α.

Single-step approach: Bonferroni

- The Bonferroni Method is the simplest way to achieve control of the FWER at any desired level α.
- Simply choose $c = \alpha/m$.

Single-step approach: Bonferroni

- The Bonferroni Method is the simplest way to achieve control of the FWER at any desired level α.
- Simply choose $c = \alpha/m$.
- With this value of c, the FWER will be no larger than α for any family of m tests.

An example

- Suppose we conduct 5 tests and obtain the following p-values for tests 1 through 5.

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

- Which null hypotheses will you reject if you wish to control the FWER at level 0.05?
- Use the Bonferroni method to answer this question.

Solution

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$p_1 = 0.042 > 0.01$
$p_2 = 0.001 \leq 0.01$
$p_3 = 0.031 > 0.01$
$p_4 = 0.014 > 0.01$
$p_5 = 0.007 \leq 0.01$

The cutoff for significance is $c = 0.05/5 = 0.01$ using the Bonferroni method. Thus we would reject the null hypothesis for tests 2 and 5 with the Bonferroni method.

How to do this in `R`?

There are several packages for multiple testing correction, as for example:

- `p.adjust` (the simplest – available at `stats` package)
- `multtest` (the most popular for gene expression data)
- `qvalue`
- `fdrtool`
- `structSSI` (for hypotheses with hierarchical or group structure)

In this example, we will apply `p.adjust`.

Read the data:

```
> raw.p.values<-c(0.042,0.001,0.031,0.014,0.007)
> raw.p.values
[1] 0.042 0.001 0.031 0.014 0.007
```

Adjust the p-values:

```
> corr.p.values.Bonf<-p.adjust(p=raw.p.values,"bonferroni")
> corr.p.values.Bonf
[1] 0.210 0.005 0.155 0.070 0.035
```

Visualize tests corresponding to the rejected null hypotheses (position of the corrected p-value in the vector):

```
> global.alpha<-0.05
> which(corr.p.values.Bonf<global.alpha)
[1] 2 5
```

Sequential Adjustments: Holm's method

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.

Sequential Adjustments: Holm's method

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest** integer k so that:

$$p_{(i)} \leq \frac{\alpha}{m-i+1} \text{ for all } i = 1, \ldots, k.$$

Sequential Adjustments: Holm's method

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest** integer k so that:

$$p_{(i)} \leq \frac{\alpha}{m-i+1} \text{ for all } i = 1, \ldots, k.$$

- If no such k exists, set $c = 0$ (declare nothing significant).

Sequential Adjustments: Holm's method

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest** integer k so that:

$$p_{(i)} \leq \frac{\alpha}{m-i+1} \text{ for all } i = 1, \ldots, k.$$

- If no such k exists, set $c = 0$ (declare nothing significant).
- Otherwise set $c = p_{(k)}$ (reject the null hypotheses corresponding to the smallest k p-values).

Sequential Adjustments: Holm's method

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest** integer k so that:

$$p_{(i)} \leq \frac{\alpha}{m-i+1} \text{ for all } i = 1, \ldots, k.$$

- If no such k exists, set $c = 0$ (declare nothing significant).
- Otherwise set $c = p_{(k)}$ (reject the null hypotheses corresponding to the smallest k p-values).
- The point here is that we do not multiply every p_i by the same factor m.

Example (cont.)

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$p_{(1)} = 0.001 \leq 0.05/(5-1+1) = 0.01$
$p_{(2)} = 0.007 \leq 0.05/(5-2+1) = 0.0125$
$p_{(3)} = 0.014 \leq 0.05/(5-3+1) = 0.0167$
$p_{(4)} = 0.031 > 0.05/(5-4+1) = 0.025$
$p_{(5)} = 0.042 \leq 0.05/(5-5+1) = 0.05$

These calculations indicate that **Holm's method** would reject null hypotheses for tests 2, 4 and 5. Here, $k = 3$ and $c = p_{(3)} = 0.014$.

How to do this in R?

Adjust the p-values:

```
> corr.p.values.Holm<-p.adjust(p=raw.p.values,"holm")
> corr.p.values.Holm
[1] 0.062 0.005 0.062 0.042 0.028
```

Visualize tests corresponding to the rejected null hypotheses (position of the corrected p-value in the vector):

```
> which(corr.p.values.Holm<global.alpha)
[1] 2 4 5
```

Some considerations

- FWER is appropriate when you want to guard against ANY false positives.
- However, in many cases (particularly in genomics) we can live with a certain number of false positives.
- FWER criteria may be too restrictive because control of false positives implies a considerable increase of false negatives.
- FWER is to conservative because it depends on the overall number of tests (m).
- Holm's method is less conservative than the Bonferroni method.
- The methods will provide the same results for many data sets, but sometimes Holm's method will result in more rejected null hypotheses.

FDR - False Discovery Rate

In practice, however, many biologists seem willing to accept that some errors will occur, as long as this allows findings to be made. For example a researcher might consider acceptable a small proportion of errors (say 10%, 20%) between his findings. In this case, the researcher is expressing interest in controlling the **false discovery rate** (FDR).

- FDR which is the proportion of false positives among all the genes initially identified as being differentially expressed.
- Unlike a significance level which is determined before looking at the data, FDR is a post data measure of confidence.
- FDR uses information available in the data to estimate the proportion of false positive results that have occurred.
- If one obtains a list of differentially expressed genes where the FDR is controlled at, say, 20%, one will expect that a 20% of these genes will represent false positive results.

	Rejected H_0	Not Rejected H_0	
True H_0	V	U	m_0
False H_0	S	T	m_1
	R	$m - R$	m

V: false positives (type I error)
T: false negatives (type II error)

- FDR is designed to control the proportion of false positives among the set of rejected hypothesis (R)
- $FDR = \frac{E(V)}{R}$

The Benjamini and Hochberg Procedure for Strongly Controlling FDR at Level α

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.

The Benjamini and Hochberg Procedure for Strongly Controlling FDR at Level α

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest integer** k so that $p_{(k)} \leq \frac{k \times \alpha}{m}$.

The Benjamini and Hochberg Procedure for Strongly Controlling FDR at Level α

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest integer** k so that $p_{(k)} \leq \frac{k \times \alpha}{m}$.
- If no such k exists, set $c = 0$ (declare nothing significant).

The Benjamini and Hochberg Procedure for Strongly Controlling FDR at Level α

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest integer** k so that $p_{(k)} \leq \frac{k \times \alpha}{m}$.
- If no such k exists, set $c = 0$ (declare nothing significant).
- Otherwise set $c = p_{(k)}$ (reject the null hypotheses corresponding to the smallest k p-values).

Example (cont.):

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$p_{(1)} = 0.001 \leq 1 \times 0.05/5 = 0.01$
$p_{(2)} = 0.007 \leq 2 \times 0.05/5 = 0.02$
$p_{(3)} = 0.014 \leq 3 \times 0.05/5 = 0.03$
$p_{(4)} = 0.031 \leq 4 \times 0.05/5 = 0.04$
$p_{(5)} = 0.042 \leq 5 \times 0.05/5 = 0.05$

The **Benjamini and Hochberg's Method** would reject the null hypotheses for all 5 tests. Here, $k = 5$ and $c = p_{(5)} = 0.042$.

How to do this in R?

Adjust the p-values:

```
> corr.p.values.BH<-p.adjust(p=raw.p.values,"BH")
> corr.p.values.BH
[1] 0.04200000 0.00500000 0.03875000 0.02333333 0.01750000
```

Visualize tests corresponding to the rejected null hypotheses (position of the corrected p-value in the vector):

```
> which(corr.p.values.BH<global.alpha)
[1] 1 2 3 4 5
```

- This correction is the least stringent of all previous options, and therefore tolerates more false positives.
- There will be also less false negative genes.
- The correction becomes more stringent as the p-value decreases, similarly as the Bonferroni step-down correction.

Some final considerations

FWER vs FDR

- The decision of controlling FDR or FWER depends on the goals of the experiment.
- If the objective is *gene fishing*, allowing a certain number of false positives to be reasonable, then FDR is preferable.
- If instead one is working with a shorter number of hypotheses, in which we want to verify if some specific ones are significant, then FWER is the appropriate criteria.
- FDRs are more appropriate in large sets of hypotheses.

Remarks

- Which multiple tests correction should be used? As long as the conditions you have for the data meet with the assumptions in particular multiple tests corrections, use the one that gives the highest power. **Using an FDR method is common these days.**
- 5% (or 95% confidence) is a convention, not a magic number (same to 10% or 1%). If you do not have any particular reason to favour a particular threshold, use a convention.

Acknowledgements:

Antónia Turkman (DEIO – FCUL) and Carina Silva-Fortes (ESTeSL – IPL), for allowing the use of some material produced by all of us in previous courses.

Bibliography:

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* Vol. 57, 289–300.

Keen, K. J. (2010). *Graphics for Statistics and Data Analysis with R*. Chapman & Hall.

Lumley, T. (2010). *Complex Surveys: a Guide to Analysis Using R*. Wiley.

Maindonald, J. (2010). *Data Analysis and Graphics Using R: an Example-Based Approach*. Cambridge University Press.

Storey, J.D. and Tibshirani, R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Preprint.