Bioinformatics

Statistical Methods in Bioinformatics

Manuela Neves

ISA/ULisboa

20 & 22 April 2020

BLOCK I

Statistical revisions using the R.
Hypothesis testing
Multiple testing. *P-values*

Plan of classes

1. References

2. The comparison of sequences
   - Alignment of pairs of sequences

3. Brief review of statistical concepts
   - Hypothesis testing
   - Functions in the R for models of random variables
   - *P-Values*

4. Contingency tables and chi-square tests
   - Independence Tests
   - Homogeneity Tests

References

- **W. Ewens and G. Grant. (2001)**. *Statistical Methods in Bioinformatics. An introduction*. Statistics for Biology and Health. Springer
- **W. P. Krijnen (2009)**. *Applied Statistics for Bioinformatics using R*. Disponível *online*
- **M. Manuela Neves (2017)**. *Introdução à Estatística e à Probabilidade com utilização do* R. *ISAPress*.
- **D. D. Pestana e S. F. Velosa (2008)**. *Introdução à Probabilidade e à Estatística.* Fundação Calouste Gulbenkian.
- **K. Seefeld (2007)**. *Statistics using R with Biological Examples*. University of New Hampshire Department of Mathematics & Statistics.

The comparison of sequences

Objective: To identify similarities/differences between DNA, RNA and Protein sequences.

When comparing sequences it is intended to

1. Analyze two or more sequences;
2. Identify differences

To do this, Sequences Alignment is performed to

1. Measure similarity between two or more sequences
2. Inferring evolutionary relationships
3. Observe conservation and variability standards for structural and functional predictions

Alignment of pairs of sequences

- Comparing two biological sequences is the same as comparing two “ strings ”.
- You have seen that there are several methods for comparing strings.
- From a biological point of view, it is possible that similarity occurs due to chance.
- The alignment is intended to identify homologous sequences.
- What we are going to deal with in these two classes is to refer to statistical procedures that allow you to interpret and understand what is going on.

Sequence comparison– hypothesis testing

The statistical procedures for comparing two sequences start with the formulation of a **null hypothesis, H_0**, for example:

H_0: given a pair of aligned amino acids, the two amino acids were generated by independent mechanisms

i.e., a little more formally, if the amino acid j occurs in any position, in the 1 ^{a} sequence, with probability p_j and the amino acid k occurs in any position, in the 2^{a} sequence, with probability p_k,
the probability of the (j, k) pair occurring in a given alignment is $p_j p_k$.

Brief review of statistical concepts

The theory of Hypothesis Tests, requires the formulation of **an alternative hypothesis.** In the case under study it is usual to consider

$\boldsymbol{H_1}$: the pair probability $(j, k) = q(j, k)$

Hypothesis testing

Formulate a **null hypothesis** *versus* an **alternative hypothesis.**

When carrying out na hypothesis test, we can say that it is necessary to execute **five steps** (with **four** before the collection, or use, of the data)

Hypothesis testing

Let's illustrate using two DNA sequences, perhaps from two different species. The | indicates that the nucleotide pairs are the same in both sequences

```
g g a g a c t g t a g a c a g c t a a t g c t a t a
|   |     |     | | |   | |                 | | |
g a a c g c c c t a g c c a c g a g c c c t t a t c
```

We have an *example of sequence-matching*

Do the two sequences have a significantly greater similarity than would be expected in two arbitrary DNA sequences, of that species, to be able to say that there is evidence that they come from a common ascendant?

Hypothesis testing

If the sequences were generated at random, the 4 letters **a, g, c,** and **t** were equally likely to occur in any position, so sequences would tend to have the same letter in about 1/4 of the positions
But ... in the 26 positions there are 11 common

How **unlikely** would this happen, if they were generated at random?

Enter here knowledge and properties of random sequences here.
Observing the 11 identities (*matches*) at 26, $11/26 = 0.42$, gives some information that something more than chance has occurred.

We are " estimating ", based on the data, a hypothetical unknown value, parameter, and we intend to make decisions about " how much we believe in that value ".

Steps in a hypothesis test

Step 1

Formulate the null hypotheses, H_0, and alternatively, H_1.

Brief Notes:

- The choice of null and alternative hypotheses must be made before data collection.
- The purpose of the test is to reject or not to reject H_0 using appropriate statistical procedures and using the data.
- What does it mean to say “ the null hypothesis is accepted ”? - means to say– there is no statistical evidence to reject it in favor of the alternative hypothesis.

Hypothesis testing

But the null hypothesis can be accepted because the alternative may not explain the data better enough.

So, better than saying accept ... you should say **not to reject** H_0

Hypothesis testing

Let's go back to the example

Let's choose H_0 and H_1.

$H_0 : p = 0.25$, meaning that each of the four nucleotides appears at any site with probability 0.25, regardless of the other nucleotides, so the two sequences were generated at random;

and we can specify that the alternative hypothesis is $p > 0.25$, or another value, for example, $p = 0.35$, as could happen if they were related.

Hypothesis testing

So in our example, it is natural to consider alternative hypothesis $p > 0.25$

Passo 2

Choice of decision error

Notes:
The decision not to reject or reject H_0 based on the data, may be incorrect.

	no rej. H_0	rej. H_0
H_0 true	correct decision	type I error,
H_0 false	type II error,	correct decision

Hypothesis testing

The errors of the decision to reject or not to reject H_0 are designated, respectively by **type I error** and **type II error,** with the probabilities associated with each of the errors commonly referred to as

$\boldsymbol{\alpha} = P$ (type I error) = P (reject $H_0|H_0$ true)
$\boldsymbol{\beta} = P$ (type II error)) = P(not rejecting $H_0|H_0$ false).

α is usually denoted as significance level of the test
and
$1 - \beta$ = P (reject $H_0|H_0$ false) test power.

Hypothesis testing

Ideal situation – have the arbitrarily small probabilities of having a Type I error and a Type II error, which is not possible to guarantee, unless the number of observations was as large as we wanted.

The dilemma is resolved, seeing that there is asymmetry in the implications of the two errors.

For example, in the example *sequence-matching*,

- there may be more concern about making a false positive claim - that the 2 sequences are similar, if there is no similarity
- and less worry in a conclusion false negative - saying there is no similarity, when there is similarity

Hypothesis testing

Usually

A value for the Type I error probability is fixed, α (very low 1 % or 5 %).

The test theory was developed to ensure that if fixed α, the Type II error has the least probability

In this Step 2 - the value of α is fixed

Hypothesis testing

Step 3

Determination of **test statistic** - is the variable that, calculated from the data, leads to decision making — leads to the acceptance or rejection of the null hypothesis

In the example *sequence-matching* a **possible test statistic** is Y - v.a. which counts the total number of *matches*.

Sometimes choosing the test statistic may not be easy !!

Hypothesis testing

Step 4

In this step, the **value of the test statistic** is determined based on the observed values.

Example with our problem

So, Y, total number of *matches* is the test statistic.

Whether the alternative hypothesis was $p = 0.35$ or $p > 0.25$ the null hypothesis $p = 0.25$ was rejected in favor of the alternative when the y observed value of Y is large enough, i.e., it is greater than some significance value K.

Hypothesis testing

If Type I error is chosen equal to 5 %, K is such that

$$\text{Prob (null hypothesis to be rejected | true)=}$$
$$Prob(Y > K|p = 0.25) = 0.05$$

If we are working with discrete variables, it may not be possible to find a K value that gives exactly that Type I error value

For the calculation of K and other amounts of interest in random variable models, let's remember the facilities of R

Functions in the R for models of random variables

- **d**function $(x, ...)$ - allows obtaining the mass probability function (discrete model) or the density function (continuous model) in x;
- **p**function$(q, ...)$ - allows obtaining the cumulative distribution function, i.e., returns the probability that the variable is less than or equal to q;
- **q**function $(p, ...)$ - allows calculating the quantile associated with the p probability;
- **r**function $(n, ...)$ - allows you to generate a sample of n pseudo-random numbers from the specified model.

Meaning:

density, probability, quantile, random

Hypothesis testing

In the case of our example, verify that

$Prob(Y > 10|p = 0.25) = 0.0400845$ e
$Prob(Y > 9|p = 0.25) = 0.09085561$
So the choice of K is made conservatively, i.e., it should be considered **$K = 10$**

Check that if, for example, you had $n = 100$, $\alpha = 0.05$ e $p = 0.25$
$Prob(Y > 31|p = 0.25) = 0.069$ e
$Prob(Y > 32|p = 0.25) = .044$
We use conservative value 32 for K.

Remark

Check that using the command `qbinom (0.95, n, 0.25)` allows you to obtain the value of K - probability quantile 0.95

Hypothesis testing

That difficulty occurs when the test statistic is a discrete r.v.

In very long sequences, binomial approximation by normal distribution can be used.

Example: $n = 1000000$ e $\alpha = 0.05$.
K can be determined by considering

$$Prob[X \geq K + 1/2] = 0.05$$

where $X \sim \mathcal{N}(\mu, \sigma)$, com $\mu = 1000000 \times 0.25 = 250000$ e $\sigma^2 = 1000000 \times 0.25 \times 0.75 = 187500$, considering continuity correction.

You get $K = 250711.74$ in practice you can use the conservative value $K = 250712$

Hypothesis testing

Step 5

Finally at this stage we will use the data !!!

Now the value of the test statistic is determined and it is checked whether is equal to or more extreme than the calculated “ point of significance ”.

The null hypothesis is rejected if the calculated value is greater than K. Otherwise (accepted) it is not rejected H_0.

P-Values

A test procedure equivalent to the one described is based on the calculation of the so-called *P-value* of the value found.

Step 4 is no longer calculated, instead ... from the data, the probability of obtaining an equal or more extreme value than that observed for the test statistic is calculated, under H_0

This probability is called *P-value*.
If *P-value* $\leq$ Type I error probability — the null hypothesis is rejected; otherwise it is not rejected

P-Values

Example

Given the null hypothesis $H_0 : p = 0.25$

What is the probability of observing 11 or more *matches* in a sequence of length 26, (example under study)?

Being $Y \frown Binomial(26, p)$ we have $P[Y \geq 11|p = 0.25] \approx 0.04$

This is the*P-value* associated to the observed value 11.

For example, if $n = 1,000$ and 278 *matches* were found, the P-value can be determined using the binomial approximation by normal as

$$Prob(X \geq 277.5)$$

P-Values

Calculation of *P-value* in the case of a bilateral alternative test

Example We want to test whether a currency is balanced. We launched 100 times and found that, for example, the " coin " side went 58 times.
The *P-value* is the probability of getting 58 or more or 42 or less given that for a bilateral alternative we have to consider the values more extremes for both tails.

Exercise: Calculate the *P-value* associated with this experiment

P-Values

Calculation of *P-value* in the case of a bilateral alternative test

The example just treated is a particular case of calculating the textit P-value, when the distribution of the test statistic, in this case, is symmetric.

In the general case, for **bilateral tests**, we adopt:

– T being the test statistic and t_{obs} the statistic value, under the hypothesis H_0, for the observed data, o *p-value* of the test is like this calculated:

- $2P[T < t_{obs}|H_0]$ se t_{obs} is reduced;
- $2P[T > t_{obs}|H_0]$ se t_{obs} is high.

(t_{obs} is reduced (high) if the estimate obtained for the parameter to be tested is lower (higher) than the value specified in H_0)

The chi-square tests

Since we are talking about hypothesis testing, we will mention some very important tests in your applications

Chi-square tests on contingency tables

Contingency tables

Suppose that individuals in a sample are classified according to two criteria (factors) A and B (qualitative or quantitative).
Consider ***r*** levels of the A criterion and ***c*** B criteria levels. Therefore the **n** observed values are classified according to 2 different factors (criteria).
Usually a table as below represents the observed frequencies o_{ij} in cell (i, j). This table is denoted a **contingency table**

	B_1	$\cdots$	B_j	$\cdots$	B_c	
A_1	o_{11}	$\cdots$	o_{1j}	$\cdots$	o_{1c}	$o_{1.}$
A_2	o_{21}	$\cdots$	o_{2j}	$\cdots$	o_{2c}	$o_{2.}$
.	.	.	.	.	.	.
A_r	o_{r1}	$\cdots$	o_{rj}	$\cdots$	o_{rc}	$o_{r.}$
	$o_{.1}$	$\cdots$	$o_{.j}$	$\cdots$	$o_{.c}$	

$\sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij} = n$ e o_{ij} represents the number of sample elements classified in the categories A_i e B_j.

Independence Tests

If the contingency table resulted from the classification of **n** individuals in the sample according to the levels of each of the criteria, as a general rule, this study intends to infer from the eventual existence of any relationship or association between the two classification criteria . The hypotheses to be tested are:

H$_{0:}$ *A* e *B* are independent *vs* **H**$_{1:}$ *A* e *B* are not independent

The test statistic is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}},$$

where e_{ij} represents the expected frequency estimate, if the H_0 hypothesis were true, i.e. $e_{ij} = \frac{o_{i.}o_{.j}}{n}$

If H_0 is true, $\boldsymbol{X^2} \sim \chi^2_{\boldsymbol{(r-1)(c-1)}}$.

The null hypothesis H_0 is rejected if $X^2_{cal} > \chi^2_{\alpha,(r-1)(c-1)}$

Exercise

An experiment was carried out to verify the effectiveness of a new flu vaccine, which was administered in a small community. The vaccine was free and had to be administered in two doses, separated by a period of two weeks. Not everyone appeared for the vaccination and some who took the 1st dose, did not come up to receive the 2nd dose. The following spring, the following information was collected on 1000 of the inhabitants of the said community:

	Not vaccinated	Uma dose	Duas doses
Gripe	24	9	13
Não gripe	289	100	565

Based on the results, check if there is sufficient evidence to indicate an association between vaccine administration and whether or not flu has occurred.

Exercise Resolution in R

We intend to test the null hypothesis, that there is no relationship between the occurrence of flu and the administration, i.e., we intend to test the hypothesis that are independent.

$H_0 : p_{ij} = p_{i.}p_{.j}, \ \forall(i,j)$
v.s.
$H_1 : p_{ij} \neq p_{i.}p_{.j}$, for at least 2 pairs (i,j)

```
gripe<-matrix(c(24,9,13,289,100,565),nc=3,byrow=T,
  dimnames=list(c("Gripe", "Nao.Gripe"),
  c("Nao.Vac.", "1Dose","2Doses")))
gripe
margin.table(gripe,1)
margin.table(gripe,2)
chisq.test(gripe)
chisq.test(gripe)$expected
chisq.test(gripe)$residuals^2
```

Remarks

Assumptions to check:

- the expected frequencies in each class should not be less than 5, when the total number of observations is ≤ 20;
- if $n > 20$ should be no more than 20 % of cells with expected frequencies below 5, nor should there be any with expected frequency less than 1.
- if in the previous cases the conditions are not met, you should add rows or columns (as long as such a junction has meaning).
- the performance of an independence test should not end with the rejection of the null hypothesis. The contribution of each cell to the value of X^2 must be analyzed.

Contingency tables - Homogeneity Tests

In the contingency tables referred to above, it was considered that the *n* dimension sample was classified according to each of the criteria, ie, the number of observations that was counted in each cell was determined after obtained the sample. Therefore, the total number of rows and columns is not under the control of the investigator. The contingency table is said to have free margins, since the totals of the margins result from the classification process. The test performed is called chi-square test of independence.

However, the total of the rows or columns of a contingency table may be under the control of the investigator, i.e., one of the margins of the table will be fixed. In this situation the test to be carried out is said to be a chi-square test of homogeneity.

Homogeneity Tests

Example - It is intended to carry out a study to find out if the behavior of drivers in the face of car accidents is different depending on the stretch of road.
In several age groups, a sample of drivers was collected and asked if they had had an accident in the previous year and, if so, whether it had been of a greater or lesser severity. The results are found in the following table:

Idade	Tipo de acidente			
	Nenhum	menor	maior	Total
Inferior a 18	67	10	5	82
18-25	42	6	5	53
26-40	75	8	4	87
40-65	56	4	6	66
mais de 65	57	15	1	73

Homogeneity Tests

Is there a difference in the distribution of responses in each age group?

So now we are concerned with answering the question: the percentage (the proportion) of accidents of each type is the same between the different age groups, ie, do the age groups show the same behavior in relation to the type of accident?

In this example, the lines represent the subpopulations from which the samples were taken. Each element of the sample was then classified into each of the three criteria: No accident, Minor accident and Major accident.

Homogeneity Tests

The test statistic is the same as in an independence test.

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}},$$

where e_{ij} is an estimate of the expected frequency.
If the populations are homogeneous, ie, if the behavior towards each type of accident is the same in each age group, then the proportion of elements in each category modality is " the same " from subpopulation to subpopulation, for example, if regarding the modality " No accidents " if you have $67/82 \approx 42/53 \approx ... \approx 297/361$, etc.

Homogeneity Tests

So we hope to find $o_{.1} o_1 / n$ observations in the 1 ª cell, then ...
The expected frequencies are then:

$$\boldsymbol{e_{ij} = \frac{o_{i.} o_{.j}}{n}}$$

If H_0 is true the test statistic, X^2, has asymptotically distribution Chi-square with $\mathbf{(r-1)(c-1)}$ degrees of freedom.

We reject the H_0 hypothesis if the calculated value, $X^2_{cal} > \chi^2_{\alpha,(r-1)(c-1)}$

Note that the test performs the same way as the independence test

Solve the exercise using the R

Independence and Homogeneity Tests

Concluding notes:

Pearson's chi-square test is a statistical test applied to categorical data, or data organized into classes and their frequencies available

A test of goodness of fit, which we will see later, establishes whether an observed frequency distribution differs from a theoretical distribution.

An independence test assesses whether observations of two variables, expressed in a contingency table, are independent of each other. A sample of dimension *n* is collected and the individuals belonging to a category (a class) of one variable and to a category(a class) of the other variable are counted.

Independence and Homogeneity Tests

A homogeneity test when we want to check if the distribution of a variable (categorical or available in classes) is the same in different populations.

A sample of size n_1 in the first population, size n_2 in the second population, etc. are collectecd and the number of individuals from each population will be counted in each value of the categorical (or the classified) variable.