**CHAPTER 6**

**Essence of Statistical Inference in Bioinformatics**

Many high-throughput experimental techniques have been developed in the past decades, including sequencing, expression arrays, mass spectrometry and many more. Statistical science provides a rigorous way to systematically retrieve information and draw conclusions from such large scale data. In this chapter, we will start from two major statistical inference tools: hypothesis testing and parameter estimation. We will then discuss considerations in statistical inference with special emphasis on large-scale genomic data. Since the data structure and biological questions to be asked greatly vary in different cases, we will focus to demonstrate the fundamental philosophy and concepts by examples in the limited paragraph. The understanding of statistical philosophy is the key to develop novel methods when facing new bioinformatic questions and performing analysis to new technology. This chapter is not meant as a statistical textbook or encyclopaedia. Interested readers can refer to any biostatistics textbook for details (e.g. Pagano and Gauvreau, 2000; Rosner, 2005).

* 1. **Hypothesis testing**

In data analyses, hypotheses like “Is this gene differentially expressed between wild type and knock-out groups of mice” or “???” naturally arise all the time. Statistical hypothesis testing is a powerful statistical tool to rigorously validate a hypothesis and make a statistical decision from experimental data. When the observed data provide hints to some biological phenomenon or signal, hypothesis testing is commonly used to decide whether such an evidence is statistically significant or merely caused by the nature of randomness. There are five major steps normally involved in hypothesis testing:

1. Quantify biological question and construct hypothesis setting: The first critical and difficult issue is how to quantify the biological question and construct a rigorous hypothesis setting. It consists of a null hypothesis (*H0*) that states “no” biological signal and an alternative hypothesis (*HA*) that states existence of the biological signal. Normally we expect to reject null hypothesis so the alternative hypothesis is “validated” statistically.
2. Design test statistic: A statistic from observed samples has to be developed in a way that relates to the hypothesis. It is a summary of data (data reduction) to indicate the degree of validity of null or alternative hypothesis.
3. Generate null distribution: After the test statistic is decided in HS.2, we need to obtain the null distribution, meaning the probability distribution of the test statistic under the null hypothesis. This step is very important to decide when to reject the null hypothesis with quantitative probability control in the next step (HS.4). Traditionally statisticians had to construct the hypothesis setting in a simpler parametric model (HS.1) and design a test statistic (HS.2) in a way that its null distribution in HS.3 (especially the cumulative distribution function) could be analytically calculated. Such an approach was a “must” before computers were developed faster and more statistical computing methods were available. Some serious drawbacks for this approach, however, are that the parametric probability distribution chosen may not fit well to the observed data and that calculation of null distribution requires heavy probability calculations and approximation theories, only achievable by a mathematically well-trained statistician. As the computing capacity greatly increased in the past two decades, more modern non-parametric approaches have been developed and the null distributions can be more affordably simulated by statistical computing.
4. Establish critical (rejection) region: After the null distribution is established, a critical region of the test statistic has to be determined so we know when to reject the null hypothesis. Intuitively the critical region is the domain of the test statistic that is the most against the null hypothesis. The threshold for the critical region is chosen quantitatively such that the probability of falling in the critical region under null hypothesis equals a pre-fixed number, called type-I error (usually denoted as α). A widely accepted α is 0.05, although in some applications α=0.01 is used to be more conservative.
5. Make decision: After HS,1-HS.4, we will declare rejection of null hypothesis if the observed statistic falls in the critical region. Otherwise we claim that we fail to reject the null hypothesis. When the null hypothesis is rejected, we statistically validate the biological signal observed for the alternative hypothesis. However, when the null hypothesis is not rejected, we cannot claim that the null hypothesis is “proved” or “validated”. It is still possible that the alternative hypothesis is true but our data do not have large enough sample size to detect it.

We should note that, in validating biological hypothesis, biologists often design experiments with treatments and with controls to ‘approve’ the hypothesis in the comparison. For statistical hypothesis testing, it is somewhat on the opposite. The key idea is to argue that the ‘exciting’ signals or patterns are not observed by chance. We normally set up the null hypothesis as no signal/pattern and try to ask: “*Under the null hypothesis, what is the chance (probability) that we observe such an exciting signal or more extreme ones?*”. If the answer to this probability (usually called p-value) is very small (the conventional threshold is 0.05), we decide that this is an unusual signal and is not observed by chance. Under this logic, we will see the statistical hypothesis testing almost always seeks to establish a null hypothesis and expect the observations to reject the null hypothesis. Another important characteristic of hypothesis testing is that failure to reject null hypothesis does not justify or accept null hypothesis. This can mean two possibilities: either null hypothesis is really true or we do not have enough sample size and statistical power to reject the null hypothesis. However, if the sample size in the problem is already very large and we still cannot reject the null hypothesis, we tend to believe the null hypothesis may likely be true or the effect size in the alternative hypothesis may be very small and ignorable.

Looking at HS.4 and HS.5 carefully, we notice that every time when we alter the significance level α, we change the corresponding rejection region and the final decision has to be re-evaluated. This inspires a natural definition of p-value for convenience in the following:

HS.4A. Calculate p-value: After the null distribution is established, we often calculate the p-value instead of the critical region. “p-value” is defined as the probability to obtain a test statistic, under the null distribution, to be at least as extreme (against the null hypothesis) as the observed test statistic. With p-value, the hypothesis testing decision can be easily made by checking if p-value is smaller than α (reject) or not. Thus alteration of α does not require reconstruction of the corresponding critical region.

Below we will walk through two examples to illustrate the concepts discussed above:

**Example 1: A classical two-sample comparison**

Imagine your collaborator performed quantitative real-time PCR to measure the expression level of a putative biomarker in 5 normal patients and 10 tumor patients. The results obtained are (*x1, x2, x3, x4, x5*)=(3.42, 2.86, 2.82, 3.04, 4.13) for normal samples and (*y1, y2, y3, y4, y5, y6, y7, y8*) =(4.42, 4.66, 5.25, 4.58, 3.79, 3.50, 3.44, 3.97) for tumors. The goal is to conclude whether this gene is a good predictive biomarker to differentiate or diagnose the disease status. A standard parametric treatment to this question from any textbook is as the following. Firstly, you assume observations of normal samples and tumor samples are from Gaussian distributions: . The question of whether this gene is a predictive biomarker can be quantified as a rigorous hypothesis setting:  versus . This completes HS.1. Next we usually use the famous t-statistic, , to quantify the degree of expression difference between the two groups (HS.2). From a classical work, we know that, under the null hypothesis, the t-statistic follows a t-distribution (degree of freedom equals the total sample size minus one) where its density and cumulative distribution function can be easily calculated (HS3). Next, we decide that, based on the question asked here, the positive large and negative large t-statistic reflects the tendency that the two sample groups have different expression levels (against the null hypothesis). The rejection region is usually set up as , where the critical value  is chosen such that type I error (false positive) is controlled at α (i.e. ) (HS4). Finally, we check whether our observed *t* falls in the rejection region *R*. If so, the null hypothesis is rejected and we conclude that this gene is a predictive biomarker. If not, we fail to reject the null hypothesis. But notice we cannot conclude this gene is not a predictive biomarker in this case. It is possible that this gene is predictive with small effect size and our sample size is too small to detect it.

 Assumptions in t-test. We will come back to this example when we discuss issues of model fitting and parametric versue non-parametric approaches.

**Example 2: Enriched CpG island in BRCA1**

CG is the least frequent dinucleotide because C in CG is easily methylated and has the tendency to mutate into T afterwards. However, the methylation is suppressed around genes in a genome. So, CG appears at relatively high frequency within these CG islands. So, finding the CG islands in a genome is an important problem.

CpG islands are regions of DNA where a cytosine (C) nucleotide occurs next to a guanine (G) nucleotide with high frequency. Such a region is often methylated by DNA methyltransferases in many eukaryotic organisms, resulting silencing of the gene expression. BRCA1, a well-known oncogene in breast cancer, is known to contain enriched CpG in its up-stream region. In Figure 1, BRCA1 sequence plus 2,000 bps up-stream and down-stream regions (84,819 bps) is retrieved and the histogram of CpG island appearance locations is shown. It is clear that the 2,000 bp up-stream region seems to contain more CpG islands (left area of the red dotted line). Given this observed data, how can we conclude that BRCA1 is indeed enriched in CpG island and the observed signal is not caused by chance?



Test 1: (Pearson’s chi-square test) The first standard testing procedure we can think of may be Pearson’s chi-square test. Chi-square test is a popular hypothesis test procedure for comparing whether the proportions of an event in different groups are identical in a categorical data. From the retrieved sequence, we found 111 CpG islands in the 2,000 bp up-stream region and 1,187 CpG islands in the entire 84,819 BRCA1 coding region (See the 2×2 table in Table 1). We first set up the hypothesis setting: *H0: the proportion of CpG island in the upstream region equals that in BRCA1.* versus *HA: the two proportions are different* (HS.1). Then we decide to use a statistic,  (where *Oij* are the observed counts in the 2×2 table and *Eij* are the expected counts when the null hypothesis is true), to measure the similarity of the two CpG island proportions in up-stream region or BRCA1 (HS.2). Thirdly, a classical approximation theory tells us that *X2* follows a  distribution with degree of freedom (2-1)×(2-1)=1 when under the null hypothesis that the two proportions are the same (HS.3). Next we determine the rejection region to be [*cα*,∞] where *cα* =7.81 (for α=0.05) is determined by P(>*cα*)=α because large *X2* value indicates higher tendency of dissimilar proportions and against null hypothesis (HS.4). Finally, comparing the observed statistic *X2*==255.7 to the rejection region, the null hypothesis is rejected (with high statistical significance) at α=0.05. Comparing the observed statistic (*X2*=255.7) to the null distribution (), the p-value is calculated as 2.2×10-16.

|  |  |  |  |
| --- | --- | --- | --- |
|  | # of bps inUp-stream region | # of bps inBRCA1 and down-stream region |  |
| appearance ofCpG island | A=111 | C=1,076 | 1,187 |
| not a CpG island | B=1,889 | D=81,743 | 83,632 |
|  | 2,000 | 82,819 | 84,819 |

Test 2: Another approach for testing this hypothesis is to ask whether the number of CpG islands observed in the up-stream region of BRCA1 is statistically significant or not. We can set up a loosely defined hypothesis setting as *HA: the observed 111 CpG island counts in the up-stream region is significantly higher than random.* versus *H0: the observed counts is obtained by random.* (HS.1) The test statistic to be inferred is naturally “*N*=the number of CpG islands in the up-stream region” (HS.2). To generate the null distribution of N, four possible procedures are outlined below.

Test 2A: We may generate random sequences of 2,000 bps with equal probabilities (i.e. generate sequence (X1,…,X2000) with P(Xi=A)= P(Xi=C) = P(Xi=G) = P(Xi=T) =0.25.) and repeat for B times. For each random sequence, compute the number of CpG islands. The empirical distribution of these B observations can serve as the null distribution of *N* (HS.3). The following shows the histogram of the B=1,000 observations and the observed *N*=111. The rejection region from the simulation is [142, ∞] (HS.4) and we conclude that the null hypothesis is not rejected (HS.5). The p-value is 0.931 (HS.4A).

 

Test 2B: In the BRCA1 sequence together with up-stream and down-stream regions, there are 27.9% of adenine (A), 21.1% of cytosine (C), 22.1% of guanine (G) and 28.8% of thymine (T). Instead of equal probabilities in Test 2A, we may utilize this background frequency information to generate the random sequences. The following figure shows the new histogram of the simulated null distribution. We can see that the null distribution slightly shifts to the left because the frequencies of C and G are lower than 0.25. The result becomes significant with the rejection region [108, ∞] (HS.4) and we conclude that the null hypothesis is rejected (HS.5). The p-value is 0.033 (HS.4A).



Test 2C: Instead of generating random sequences, we may randomly select subsequences from the genome to generate the null distribution. In Figure ?C, B sequences of 2,000 bp long are uniformly randomly selected from the 24 chromosomes (22 chromosomes plus X and Y sex chromosome) including both coding and non-coding regions. The numbers of CpG island appearances are calculated and the histogram is shown (HS.3). The rejection region obtained is [?, ∞] (HS.4) and we conclude that the null hypothesis is rejected (HS.5). The corresponding p-value is 0 (HS.4A).

Test 2D: We may further improve Test 2C by obtaining the 2,000 bp sequences from the up-stream regions of real genes. We first identify 18,696 Entrez genes that have exact chromosomal start and ending locations in the database. B up-stream sequences of 2,000 bp long are randomly selected to generate the null distribution in Figure ?D (HS.3). The rejection region becomes [?, ∞] (HS.4) and we conclude that the null hypothesis is rejected (HS.5). The corresponding p-value is 0 (HS.4A).

Test 3: A third approach for testing this hypothesis is to ask whether the high odds ratio of CpG island observations in the up-stream region versus the remaining gene sequence is statistically significant or not. The test statistic defined as the odds ratio is γ=(CpG ratio in the 2K bp upstream region)/ (CpG ratio in the remaining gene sequence) (HS.2). The hypothesis setting becomes *H0: the observed* γ *is significantly higher than random.* versus *HA: the observed* γ *is obtained by random.* (HS.1) To generate the null distribution of N, four corresponding procedures from Test 2 can be similarly obtained.

Test 3A: B random sequences of 84,819 bps with equal probabilities are simulated and the corresponding γ statistics are calculated. The empirical distribution of these B observations can serve as the null distribution of *N* (HS.3) in Figure ?A. The observed γ is . The rejection region from the simulation is [1.135, ∞] (HS.4) and we conclude that the null hypothesis is rejected (HS.5). The p-value is 0 (HS.4A).

Test 3B: Based on the A, C, G and T frequencies, we may simulated B random sequences of 84,819 bps and calculate the γ statistics. The histogram of null distribution is displayed in Figure ?B. The rejection region from this test is [1.161, ∞] (HS.4) and we conclude that the null hypothesis is rejected (HS.5). The p-value is 0 (HS.4A).

Test 3C: Instead of generating random sequences, we may randomly select subsequences from the genome to generate the null distribution. In Figure ?C, B sequences of 2,000 bp long are uniformly randomly selected from the 24 chromosomes (22 chromosomes plus X and Y sex chromosome) including both coding and non-coding regions. The numbers of CpG island appearances are calculated and the histogram is shown (HS.3). The rejection region obtained is [?, ∞] (HS.4) and we conclude that the null hypothesis is rejected (HS.5). The corresponding p-value is 0 (HS.4A).

Test 3D: We may further improve Test 2C by obtaining the 2,000 bp sequences from the up-stream regions of real genes. We first identify 18,696 Entrez genes that have exact chromosomal start and ending locations in the database. B up-stream sequences of 2,000 bp long are randomly selected to generate the null distribution in Figure ?D (HS.3). The rejection region becomes [?, ∞] (HS.4) and we conclude that the null hypothesis is rejected (HS.5). The corresponding p-value is 0 (HS.4A).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | p-value | computing time (s) |  |  |
| Test statistic 1: chi-square statistic |
| Test 1 | 1.5E-57 | 0.09 |  |  |
| Test statistic 2: N=number of CpG islands |
| Test 2A | 0.93 | 13.26 |  |  |
| Test 2B | 0.033 | 12.89 |  |  |
| Test 2C | 0.011 | 2113.02 |  |  |
| Test 2D | 0.067 | 2010.22 |  |  |
| Test statistics 3: T=odds ratio of CpG island appearances in upstream region versus remaining |
| Test 3A | 0 | 60.87 |  |  |
| Test 3B | 0 | 61 |  |  |
| Test 3C | 0.004 | 2123.52 |  |  |
| Test 3D | 0.047 | 1809.16 |  |  |

Three different statisticians may propose three different hypothesis testing procedures. It is not to say that the hypothesis testing is not rigorous or may be flawed by human bias (although it happens sometimes). The key is to scrutinize the five steps in hypothesis testing and see whether the assumptions and null hypotheses conform to the biological contents and whether any statistical assumption or approximation is made. Another issue to consider for hypothesis testing in large-scale genomic data is the requirement of computation time and memory (i.e. feasibility of the procedure). Let us review the four tests we introduced. The first chi-squared test