Classical Hypothesis Testing and GWAS

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

November 9, 2010

Outline

General Principles

- Classical statistical hypothesis testing involves the test of a null hypothesis against an alternative hypothesis.
- The procedure consists of five steps, the first four of which are completed before the data to be used for the test are gathered, and relate to probabilistic calculations that set up the statistical inference process.

Step 1

- The first step in a hypothesis testing procedure is to declare the relevant null hypothesis H_0 and the relevant alternative hypothesis H_1.
- The choice of null and alternative hypotheses should be made before the data are seen. To decide on a hypothesis as a result of the data is to introduce a bias into the procedure, invalidating any conclusion that might be drawn from it. Our aim is eventually to acept or to reject the null hypothesis as the result of an objective statistical procedure, using data in our decision.

"the null hypothesis is accepted"

- It is important to clarify the meaning of the expression "the null hypothesis is accepted."
- In the conservative approach to statistical hypothesis testing as outlined below, this expression means that there is no statistically significant evidence for rejecting the null hypothesis in favor of the alternative hypothesis.
- For reasons discussed below, the null hypothesis is often a particular case of the alternative hypothesis, and when it is, the alternative hypothesis must explain the data at least as well as the null hypothesis.
- Despite this, the null hypothesis might well be accepted, in the above sense, in that the alternative hypothesis might not explain the data significantly better than does the null

"in favor of the alternative hypothesis"

- It is important to note the words "in favor of the alternative hypothesis" in the above.
- Suppose that the null hypothesis is that the probability of success p in a binomial distribution is $1/2$ and the alternative is that this parameter exceeds $1/2$. Suppose further that in 1,000 trials, only 348 sucesses are observed.
- The null hypothesis is accepted in favor of the alternative since the alternative hypothesis does not explain this result significantly better than does the null hypothesis - in fact it explains it less well than does the null hypothesis.
- Nevertheless, it would be unreasonable to believe that the null hypothesis is true: the data clearly suggest that $p < \frac{1}{2}$
- Thus accepting a null hypothesis in favor of some alternative

Hypotheses: simple or composite

- A hypothesis can be *simple* or *composite*. A simple hypothesis specifies the numerical values of all unknown parameters in the probability distribution of interest. In the above example, both null and alternative hypotheses are simple. A composite alternative does not specify all numerical values of all the unknown parameters. Suppose you have a hypothesis that specifies a distribution with only one parameter p. The hypothesis "$p = 0.5$ is simple. The hypothesis"p exceeds 0.25" is composite. It is also $one - sided$ $(p > 0.25)$ as opposed to $two - sided$ $(p \neq 0.25)$.
- However, for technical reasons associated with the hypothesis testing theory, it is often advantageous to make the null hypothesis a particular case of the altnerative hypothesis. If

Unknown parameters

- Hypotheses usually involve the value of some unknown parameter (or parameters). We generically denote the parameter of interest by θ, although in some cases we use a more specific notation (such as μ for a mean). The nature of the alternative hypothesis is determined by the context of the test, in particular whether it is one-sided up (that is the unknown parameter θ exceeds some specified value θ_0), one-sided down ($\theta < \theta_0$), or two-sided ($\theta \neq \theta_0$). In many cases in bioinformatics the natural alternative is both composite and one-sided.

Type I error and Type II error

- Step 2 of the hypothesis testing procedure consists in choosing the numerical value for the Type I error.
- Since the decision to accept or reject H_0 will be made on the basis of data derived from some random process, it is possible that an incorrect decision will be made, that is,
  - to reject H_0 when it is true – a *Type I error*), or
  - to accept H_0 when it is false – a *Type II error*).
- When testing a null hypothesis against an alternative it is not possible to ensure that the probabilities of making a Type I error and a Type II error are both arbitrarily small unless we are able to make the number of observations as large as we please. In practice we are seldom able to do this.

Type I error and Type II error

- This dilemma is resolved in practice by observing that there is often an asymmetry in the implications of making the two types of error. For example, there might be more concern about making the false positive claim and less concern about making the false negative conclusion.
- For this reason, a frequently adopted procedure is to focus on the Type I error, and to fix the numerical value α of this error at some acceptably low level (usually 1% or 5%), and not to attempt to control the numerical value of the Type II error.
- The choice of the values 1% and 5% is reasonable, but is also clearly arbitrary. The choice 1% is a more conservative one than the choice 5%.

Determining the Test Statistic

- The third step in the hypothesis testing procedure consists in determining a test statistic.
- This is the quantity calculated from the data whose numerical value leads to acceptance or rejection of the null hypothesis.
- Although sometime there are reasonable choices for test statistics that are usually used, in more complicated cases the choice of a test statistic is not straightforward. The main problem is that of deriving test statistics that, for a given Type I error, minimize the probability of our making a Type II error, given the number of observations to be made.
- There is a substantial body of statistical theory associated with such an optimal choice of a test statistic.

Determining the Significance Point K

- The next step in the procedure consists in determining those observed values of the test statistic that lead to rejection of H_0.
- This choice is made so as to ensure that the test has the numerical value for the Type I error chosen in Step 2.
- Suppose that the null hypothesis is $p = 0.25$
- Suppose that the alternative hypothesis is simple e.g., "$p = 0.35$" or the alternative hypothesis is composite "$p \geq 0.25$"
- Suppose we choose as as test statistic the value observed for a numerical random variable Y whose value the larger it is the more away from the Null Hypothesis it is.
- Then the Null Hypothesis $p = 0.25$ is rejected in favor of the alternative when the observed value y of Y is sufficiently large

Determining the Significance Point K

- If for example the Type I error is chosen as 5%, K is found from the requirement

$$\text{Prob (null hypothesis is rejected when it is true)}$$
$$= Prob(Y \geq K|p = 0.25) = 0.05$$

- In practice, when discrete random variables are involved, it may be impossible to arrive at a procedure having exactly the Type I error chosen. This difficulty arises here: It is impossible to find a value of K such that the above equation is satisfied exactly.
- In practice, the choice of K is made by a conservative procedure
- This difficulty is to be taken as understood in all testing

Sufficiently large or sufficiently small

- In the above example the null hypothesis is rejected if Y is sufficiently large. If the alternative hypothesis had specified a value of p that is less than 0.25, then the null hypothesis would be rejected for sufficiently small Y.
- In many test procedures the null hypothesis does not specify the numerical values of all the parameters involved in the distribution of the random variables involved in the test procedure. In such a case problems can arise in the testing procedure since there might be no unique significance point (such as K above) having the property that the probability that the test statistic exceeds K is equal to the Type I error no matter what the values of the parameters not specified by the null hypothesis.

The final step of the hypothesis testing process

- The final step in the testing procedure is:
  - to obtain the data, and
  - to determine whether the observed value of the test statistic is equal to or more extreme than the significance point calculated in Step 4, and
  - 1. to reject the null hypothesis if it is.
    2. Otherwise the null hypothesis is accepted.

The P-Values

- A testing procedure equivalent to that just described involves the calculation of a so-called P-value, or achieved significance level.
- Here Step 4, the calcuation of the significance point such as K is not carried out.
- Instead, once the data are obtained, we calculate the null hypothesis probability of obtaining the observed value, or one more extreme, of the test statistic.
- This probability is called the P-value.
- If the P-value is *less than or equal to* the chosen Type I error, the null hypothesis is rejected.
- **This procedure always leads to a conclusion identical to that based on the significance point approach.**

beginframe

- For the test of a simple null hypothesis against a simple alternative, with a fixed number of observations, the choice of the Type I error implicitly determines the numerical value β of the Type II error.
- Equivalently it determines the power of the test, defined as the probability 1 - β of rejecting the null hypothesis when the alternative is true.
- When the alternative hypothesis is composite, the probability that the null hypothesis is rejected will normally depend on the actual value of the parameter (or parameters) concerned in the test.
- There is therefore no unique value for the Type II error, and no unique value for the power of the test, under the

alternative hypothesis.

- In this case the principle adopted in Step 3, namely that of choosing a test statistic that maximizes the power of the test, becomes more difficult to apply than in the case when the alternative hypothesis is simple.