BENG 183. Applied Genomic Technologies

Class handouts

The instructor's cheat sheet for hypothesis testing

To perform a hypothesis test, we will need the following steps:

1. Forming two competing hypotheses, called **the null (H0) and the alternative hypothesis (H1)**.
    a. Tip: The keyword is "competing". If we consider H0 and H1 as two sets, they must be disjoint.
        i. For example, please consider whether these following pairs of hypotheses can be regarded as competing hypotheses:
            1. A table surface is flat vs. a table surface is not flat.
            2. $u=0$ vs. $u>0$.
            3. $u=0$ vs. $u\geq 0$.
            4. $u=0$ vs. $u\neq 0$.
    b. Tip: put the simpler hypothesis as H0. Either of the two hypotheses can be regarded as H0, however, the procedure for testing the two hypotheses will be easier if the simpler hypothesis is designated as H0.
        i. For example, which way of formatting H0 and H1 is better:
            1. H0: $u=0$; H1: $u\neq 0$
            2. H0: $u\neq 0$; H1: $u=0$
2. Generating or getting data. Our general idea is to use the data generated from experiments to test the hypothesis, that is to argue which of the two competing hypotheses is more likely to be supported by data.
    a. Tip: the key word is "argue". We will see that the entire testing is a process of forming an argument.
    b. Tip: the basis of this argument is the data.
        i. If there are already data generated by experiments, no need to do anything.
        ii. If there are no data generated yet, do the experiments to generate the data.
        iii. The experiments must be relevant to the hypotheses, thus can be used for an argument about the hypotheses (we will revisit this point in the discussion of Test Statistic).
3. Summarizing the data into a **Test Statistic**. The data can be a large set of numbers, which can hardly be directly used for making any argument. Thus, we must summarize the data into a single number to form our argument. This single number is called a Statistic.
    a. Tip: the first keyword is a single number. For example, suppose my 3 experiments generated 3 data points, i.e. x1, x2, x3, which of the following is a Statistic:
        i. x1+x2+x3
        ii. x1-x2-x3
        iii. x1-x2

iv. x2

b. Tip: the second concept is to form an argument. To be able to form an argument, the magnitude of this Statistic must reflect the degree of support to one of the two hypotheses. When the magnitude of this Statistic reflects the degree of support to one of the two hypotheses, we call this Statistic the Test Statistic. Thus, we can make an argument based on how large of small the Test Statistic is.
    i. For example, suppose we are testing two brands of an experimental reagent, Brand X, a classical brand, and Brand Y, a new brand. for which produced a greater yield of DNA from a DNA extraction experiment. We performed the experiment with Brand X 3 times, with the DNA yield of x1, x2, x3. We performed the experiment with Brand B 4 times, with the DNA yield of y1, y2, y3, y4. What hypotheses can we formulate?
        1. Answer: H0: the average yield of Brand X equals the average yield of Brand Y. H1: the average yield of the new brand (Y) is larger than the average yield of the classical Brand (X).
    ii. Which of the following statistic (denoted as t) is a good Test Statistic? There can be more than 1 correct answer:
        1. t= x1+x2+x3
        2. t= x1+x2+x3+y1+y2+y3+y4
        3. t= (x1+x2+x3+y1+y2+y3+y4)/7
        4. t= (x1+x2+x3)-(y1+y2+y3+y4)
        5. t= [(x1+x2+x3)/3-(y1+y2+y3+y4)/4]/σ, where σ is a constant
        6. t= [(x1+x2+x3)/3]/[(y1+y2+y3+y4)/4], assuming y1+y2+y3+y4>0
    iii. Tip: there are more than one Test Statistic for testing a pair of competing hypotheses.

c. At this point, based on the magnitude of our Test Statistic (t), we can already subjectively judge which hypothesis more believable. (Please recall at the beginning of this course, we mentioned that probability can be interpreted as subjective belief.)

d. Our last question is how to quantify our subjective belief? This question leads to the introduction of the p-value.

4. Calculating **p-value**. The p-value can be thought as the probability of seeing the currently observed value of the Test Statistic (t) or *more extreme* values of this Test Statistic (T) if we were to repeat the experiments in future for many times and the null hypothesis is true:
    a. The above statement can be written as: P-value = $P(T \geq t \mid H0)$
    b. Note that we have unconsciously introduced a random variable, called the Test Statistic (T). Also note that t is the observed value of this RV based on the currently finished experiments.
    c. To calculate the p-value, we note that p-value = $1 - F_{T|H0}(t)$, where $F_{T|H0}(t)$ is the CDF of T when H0 is true. Please do not be scared by T|H0. T|H0 is just a random variable (distribution). We call this distribution the null distribution or the distribution of the Test Statistic under the null hypothesis.

    d. Finally, as long as we can obtain the CDF of the null distribution, we can calculate the p-value.
    e. How to derive the CDF of the null distribution? There are two ways.
        i. First, people can try to derive the mathematical form of the CDF under some assumptions on the distribution of the original experiment. Each of these previously derived and documented CDFs is usually titled with a name, such as the T distribution. Coupled with the name of the CDF is statistical test, such as the T test.
        ii. Second, people can use computer simulation to produce random outcomes under H0. In the Brand X vs. Brand Y example, one way to simulate random outcomes is to keep the yields but randomly swap the labels (x, y). Summarizing these simulated data points can produce an approximation to the null distribution.
5. Making a decision based on p-value. The decision is either "Reject H0" or not reject. We use p-values to quantitatively assess how much belief we have for H1, i.e. against H0. The smaller the p-value, the great belief we have against the H0. People often choose an arbitrary cutoff to p-value, such as 0.05, to make a binary judgement. For example, people often say: "Since the p-value is smaller than 0.05, we reject the null hypothesis" or "We cannot reject the null hypothesis because the p-value is greater than 0.05."
6. (Optional) Making a decision based on the **acceptance region**. This is an alternative (slightly old fashioned) way for making a decision. Since the magnitude of the observed value (t) of our Test Statistic directly relates to the degree of evidence for (or against) H0, we can use a threshold (δ) on the Test Statistic for making the decision. For example, if the larger t is the greater the evidence is against H0, then we can decide to:
    a. Reject H0, if $t > \delta$.
    b. Not to reject H0, if $t \leq \delta$.

    In this case, $(-\infty, \delta)$ is the **Acceptance Region**. The statistical test is completely defined when the Test Statistic (T) and the acceptance region are given. For a completely defined test, we can obtain the following probabilities:

    a. P(Reject H0|H0). The action (decision) of "Reject H0|H0" is called the **Type I error**, also referred to as the false positive.
    b. P(Do not reject H0|H1). The action (decision) of "Not to reject H0|H1" is called **the Type II error**, also referred to also false negative.
    c. In the above example, $P(\text{Reject H0}|\text{H0}) = P(t > \delta|\text{H0})$, $P(\text{Do not reject H0}|\text{H1}) = P(t \leq \delta|\text{H1})$.
    d. Since the CDF of T|H0 is often given, at least in those "named tests" such as the T test, $P(t > \delta|\text{H0})$ can be computed by $P(t > \delta|\text{H0}) = 1\text{-}F_{T|H0}(\delta)$. This probability is called the **significance level**, denoted as α.
    e. As long as the CDF of T|H0, denoted as $F_{T|H0}(\delta)$, is known, we can calculate α based on the acceptance region $(-\infty,\delta)$ or calculate δ based on the significance level (α). This is because $\alpha = P(t > \delta|\text{H0}) = 1\text{-}F_{T|H0}(\delta)$.

Tutorial materials

1. **Getting started with the Linux environment:** https://github.com/Irenexzwen/BIOE183/blob/master/Tutorial1_Preparation.md
2. **Working with raw RNA sequencing data:** https://github.com/Irenexzwen/BIOE183/blob/master/Tutorial2_RawData.md
3. **Mapping RNA-seq reads and quantifying gene expression:** https://github.com/Irenexzwen/BIOE183/blob/master/Tutorial3_Mapping_and_qualification.md
4. **Identifying differentially expressed genes:** https://github.com/Irenexzwen/BIOE183/blob/master/Tutorial4_DE.md
5. **A tutorial for writing markdown (.md) files:** https://github.com/Zhong-Lab-UCSD/BENG183/blob/master/markdown_tutorial.md