Multiple Comparisons

By Jake Sauter

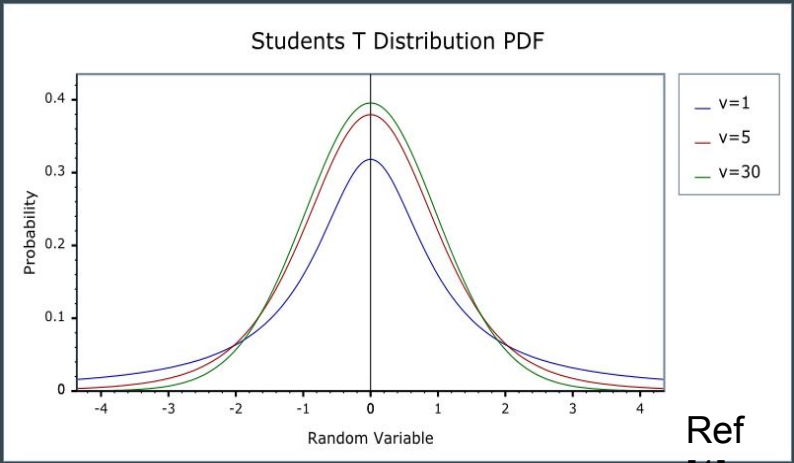
- Hypothesis Testing
- T Test
  - Application to Microarray Data
- Multiple Comparisons
  - Šidák correction
  - Bonferroni correction
  - Step-Wise correction
  - FDR correction
  - Permutation correction
  - SAM

Hypothesis Testing

- Hypothesis Test - A statistical test in which a null hypothesis (H_0) is assessed through the use of a test statistic, comparing this test statistic to a critical value
  - In most cases if the test statistic is larger than the critical value, then H_0 is rejected.
- P Value - The probability of rejecting a null hypothesis H_0 when it is true due to random chance / sampling error
- Significance Level - A significance level α is the acceptable p value such that H_0 is rejected in a statistically significant manner
- Critical Value - A quantity derived from the statistical distribution of the test statistic and the accepted p value.
- Test Statistic - A quantity derived from a sample, compared to a critical value to test for the truth value of the null hypothesis

T Test

- Student's T distribution - A statistical distribution that can be used in comparing means of normally distributed data sets in situations with unknown standard deviation σ and a small sample size. This distribution comes from dividing a Normal Distribution with a Chi-Square Distribution



Ref [1]

T Test

- The T test is used to test if there is a difference in the mean of two groups
- The test statistic for the T test, denoted T here, follows the Student's T distribution and is defined to be the following

$$T = \frac{\text{signal}}{\text{noise}} = \frac{\text{diff between group means}}{\text{variability of groups}} = \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- If the test statistic falls above the critical value on the distribution, it indicates that the means of the two groups are statistically significantly different

Applying the T-test to Microarray Data

- The T-test can be useful in microarray analysis to analyze if gene expression levels are different for a gene from a group of control individuals and group of patients
- This test is not ideal, as applying the t-test with α=.05 significance level to a list of 10,000 genes will produce approximately 500 genes that appear to be differentially expressed even if they are random
- Let's look at some simple probabilities

$$P(\text{correct}) = (1-p) \quad \text{(Eq. 1)}$$

$$P(\text{globally correct}) = (1-p) * (1-p) * \ldots * (1-p) = (1-p)^R \text{for R genes} \quad \text{(Eq. 2)}$$

$$P(\text{wrong somewhere}) = 1 - P(\text{globally correct}) = 1 - (1-p)^R \quad \text{(Eq. 3)}$$

Multiple Comparisons

- Family-Wise Error Rate (FWER) - Probability of having a Type 1 error (false positive) in any comparison in the data set
- We would like to control this overall rate of false positives, framing Eq. 2 in terms of significance levels, we arrive at

$$\alpha_e = 1 - (1 - \alpha_c) \quad \text{(Eq. 4)}$$

  where α_e is the fp rate at experiment level and α_c is the fp rate at single gene level

- Solving eq 4 for α_c results in the Šidák correction for multiple comparisons

$$\alpha_c = 1 - \sqrt[R]{1 - \alpha_e} \quad \text{(Eq. 5)}$$

Corrections for Multiple Comparisons

- Bonferroni notes that for small *p*, Eq. 4 can be approximated by taking only the first two terms of the binomial expansion of $(1-p)^R$, resulting in the Bonferroni correction for multiple comparisons

$$\alpha_c = \frac{\alpha_e}{R} \quad \text{(Eq. 6)}$$

- Unfortunately, both mentioned corrections require a very small significance level for any reasonable R value.
- These corrections are sufficient but not necessary for declaring a gene differentiable between control and patient groups

Step-Wise Correction

- The Holm's step-wise group of methods allow less conservative adjustments of the *p*-values, ordering genes in increasing order of their *p*-value and making successive smaller adjustments
- Procedure:
  - Choose the experiment-level significance α_e
  - Order the genes in the increasing order of individual *p*-values
  - Compare the *p*-values of each gene with a threshold that depend on the position of the gene in the list of ordered values. The thresholds are as follows:

    $\frac{\alpha_e}{R}$ for the first gene, $\frac{\alpha_e}{R-1}$ for the second gene, and so on
  - Let *k* be the largest *i* for which $p_i < \frac{\alpha_e}{R-i+1}$. Reject the null hypothesis for *i* = 1, 2, ... , k

False Discovery Rate (FDR)

- The FDR correction procedure allows for some dependencies between variables, while the previous methods act on the assumption of independence
- FDR correction procedure:
  - Choose the experiment-level significance α_e
  - Order the genes in the increasing order of individual *p*-values
  - Compare the *p*-values of each gene with a threshold that depends on the position of the gene in the list of ordered values. The thresholds are as follows:

    $\frac{1}{R}\alpha_e$ for the first gene, $\frac{2}{R}\alpha_e$ for the second gene, and so on
  - Let *k* be the largest *i* for which $p_i < \frac{i}{R}\alpha_e$. Reject the null hypothesis for *i* = 1, 2, ... , k

Permutation Correction

- The Westfall and Young (W-Y) step-down correction is a more general method that adjusts the *p*-value whole taking into consideration the possible correlations
- This method permutes the classes individuals thousands of times, running a T-test after every permutation
- The *p*-value for a gene *i* will the the proportion of times the value of *t* calculated for the real labels is less than or equal to the value of *t* calculated for a random permutation

$$p\text{-value for gene } i : \frac{\text{number of permutations for which } t_j^{(b)} \geq t_i}{\text{total number of permuatations}}$$

where $t_j^{(b)}$ are the calculated t-values from gene j and permutation b

Significance Analysis of Microarrays (SAM)

- Tusher et al. have reported that permutation testing was still too stringent for their microarray data. In response, they formulated their own method
- SAM assigns a score to each gene taking into consideration the relative change of each gene expression level with respect to the standard deviation of repeated measurements
- SAM uses a test statistic similar to *T*, in that it expresses the difference between means in terms of standard deviations

$$d_i = \frac{\bar{x_i}_1 - \bar{x_i}_2}{s_i + s_0}$$

SAM contd.

- SAM calculates a gene-by-gene variance which will allow for the selection of the appropriate genes independently of their expression levels
- SAM uses the same permutation idea to estimate the percentage of genes identified just by chance
- FDR in SAM:
  - Fix a threshold for differentially expressed genes
  - Count how many genes are reported as differentially expressed in each permutation
  - Calculate the median number of false positives across all permutations
  - Calculate the FDR as the number of false positives divided by the number of genes in the original data

References

[1] Maddock, John, et al. “Students t Distribution.” Boost C++ Libraries, Boost Software, 2008, www.boost.org/doc/libs/1_36_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/students_t_dist.html.

[2] Drăghici Sorin. Statistics and Data Analysis for Microarrays: Using R and Bioconductor. Chapman and Hall, 2012.