STA613/CBB540: Statistical methods in computational biology

Hypothesis Testing (1/22/13)

Lecturer: Barbara Engelhardt *Scribe: Lisa Cervia*

1. Hypothesis Testing

1.1. **Classical.**

In classical hypothesis testing we are concerned with evaluating the probability of the data given that a hypothesis is true: $P(D|H = t)$

In Bayesian hypothesis testing, we instead consider the probability that the hypothesis is true given a set of data: $P(H = t|D)$

Throughout both paradigms, we compare the null hypothesis, or one specific model of the data, with an alternative hypothesis, a second possible model of the data. We are interested in detecting data points for which the alternative model is a significantly better at explaining the data than the null model:

Null hypothesis: H_0

Alternative hypothesis: H_A

So, as an example, consider the hypothesis that, for two nucleotide sequences of length n, that they are somehow are homologous at a higher rate than expected by chance (assuming the probability of a nucleotide match in the null hypothesis is $\frac{1}{4}$.)

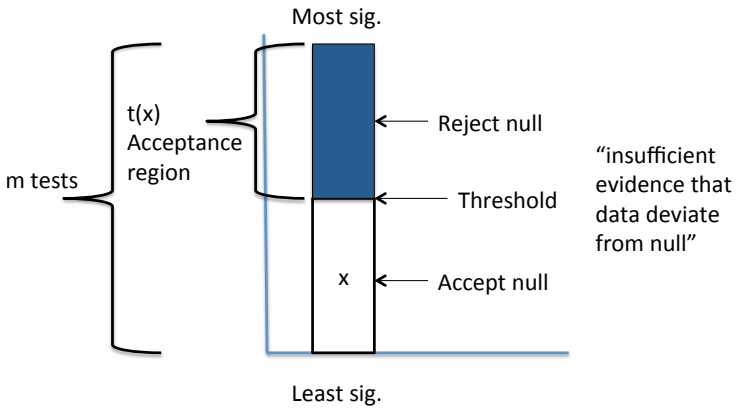
Null, $H_0 : X \sim \text{Binomial}(n, p = \frac{1}{4})$

One sided test, $H_A : X \sim \text{Binomial}(n, p < \frac{1}{4})$

Two sided test, $H_A : X \sim \text{Binomial}(n, p \neq \frac{1}{4})$.

1.2. **Test Statistic.** Let $t(x)$ be the *threshold* of a test statistic evaluated on m data points, or *features*, where $x = \{x_1, ..., x_m\}$, and let m be the number of tests that are performed. For those values that are more extreme than the threshold $t(x)$, we are going to call those feature *significant*, and for them we will say that we reject the null hypothesis in favor of the alternative hypothesis. Alternatively, for those values of the statistic that are less extreme than the threshold, we will call those features *insignificant* (Figure 1), where there is insufficient evidence that the data deviate from the null hypothesis.

In our original data set x, there will be both *truly null* (those generated from the null model and *truly alternative* (those generated from the alternative model) features. Our hope is that the statistical test we are performing will identify all of the *truly alternative* features as significant (and none of the truly null features). There are four possible ways this can turn out (Figure 1):



- true positive (TP): truly alternative features that are called significant
- false positive (FP): truly null features that are called significant (also called type I error)
- true negative (TN): truly null features that are called insignificant
- false negative (FN): truly alternative features that are called insignificant (also called type II error).

	H_0=T	H_0=F
Accept H_0	TN	FN
Reject H_0	FP	TP

Let α=Pr(null is rejected when H_0 is true), or the probability of a false positive. α is called the *significance level*, and is often taken to be 0.05. In this hypothesis testing framework, we will control the value α explicitly, and choose the threshold $t(x)$ based on some value $\alpha = 0.05$ (for example).

1.3. **Statistical power.**

Statistical power (1 - β): probability of appropriately rejecting the null, or probability of not committing a type I error.

β = Pr(null is accepted when it is false), then $1 - \beta$ = Pr(null is rejected when false).

Returning to our example of nucleotide sequence matching:

$$H_0 : X \sim Binomial\left(n, p = \frac{1}{4}\right)$$

$$H_A : X \sim Binomial\left(n, p > \frac{1}{4}\right)$$

where $Y = 32, 33$ (the number of matches equals Y) and $n = 100$. Our statistic will be the probability of finding this number of matches under the null hypothesis:

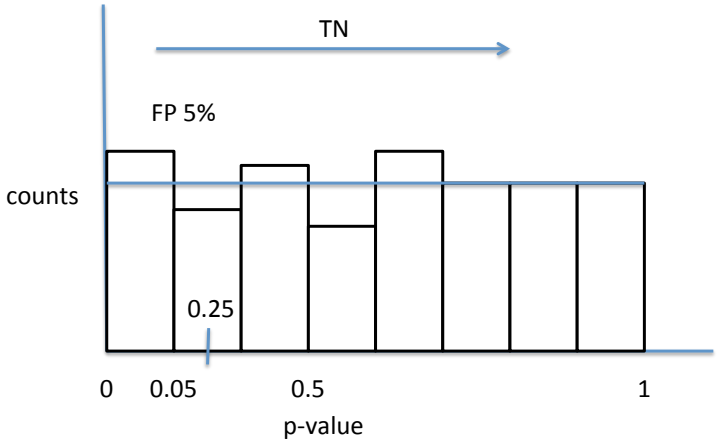
$P(Y \geq 32 | p = 0.25, n = 100) = 0.069$

$P(Y \geq 33 | p = 0.25, n = 100) = 0.044$

So if there are 33 or more matches, and your significance level is $\alpha = 0.05$, you reject the null.

P-value: the probability of obtaining an observed value under the null hypothesis distribution that is more extreme than the statistic computed on this feature.

So, in our example, the p-value for $y = 32$ is 0.069.



For a set of features generated from the null hypothesis, the corresponding p-values are expected to have a uniform distribution. Intuitively, consider that the probability at p-value = 0.5 of finding a smaller p-value at random in this null sample is exactly 0.5, and similarly for p-value = 0.25, etc.

2. Paper: "Statistical significance for genomewide studies" [Storey & Tibshirani, 2003]

2.1. **False Positive Rate.** False positive rate: given a rule for calling features significant, FPR is the rate that truly null features are called significant:

$$E[FPR] = \frac{FP}{FP + TN}.$$

- 6215 genes tested (m)
- p-value cutoff $t(x) = 0.0085$ (false positive rate)

Under the null, you would expect 53 FP by chance:

$$E[|FP|] = m \cdot FPR$$

$$53 = 6215 \cdot 0.0085$$

This does not consider how many truly alternative features you expect in your data. If you expect 100 truly alternative features, and find all of them significant, then in the set of significant features, more than 1/3 of them will be false positives.

Let m_0 = number of truly null features and $m_1 = m - m_0$ = number of truly alternative features.

2.2. **Family-wise Error Rate.** Family-wise error rate bounds the probability of more than 1 FP expected in our significant set of features by our significance level. Assuming all of our tests are independent:

$P[FP \geq 1] \leq \alpha$

In order to get this, from our definition of false positive rate above, we can scale our threshold $t(x)$:

$$t(x) \leq \frac{\alpha}{m}$$

$$t(x) \leq \frac{0.05}{6215}$$

In particular, the *Bonferroni correction* scales the significance level α by the total number of tests. This is a very conservative correction; in practice, the tests will not often be independent.

Sensitivity and specificity are important to understand in this context:

- Sensitivity: proportion of true positives: sensitivity $= \frac{TP}{TP + FN}$
- Specificity: proportion of false negatives: specificity $= \frac{TN}{TN + FP}$

FPR = 1 - specificity.

2.3. **False Discovery Rate.** Perhaps more intuitive is the false discovery rate.

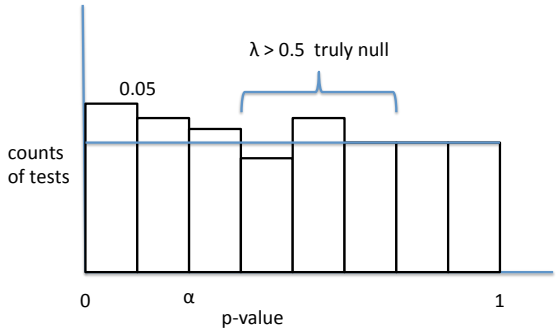
False Discovery Rate (FDR): rate that significant features are truly null

FDR bounds the proportion of features that were found to be significant that are false positives.

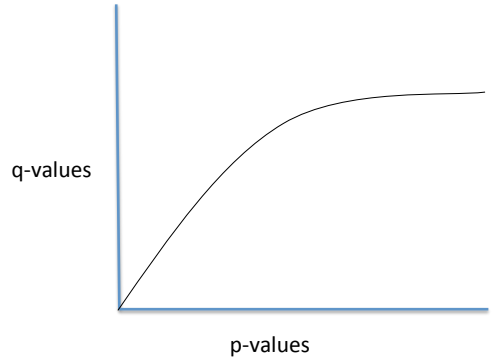
$$\text{E[FDR]} = \frac{FP}{FP + TP}$$

where $FP + TP > 0$ (i.e., there is at least one significant features).

An FDR = 5% means that, of the set of significant features, one out of 20 of them (or 5%) are expected to be false positives.



FDR is bounded by the q-value, which quantifies the probability of this feature being a false positive, given that it is called significant. There is a monotone relationship of p-values and q-values, and much of this paper discusses how to derive a set of q-values given a set of p-values (Figure 3 for intuition).



As with FPR, FDR assumes the independence of tests. For q-value $\leq t(x)$, then, we can say that the FDR $\leq t(x)$.

2.4. **Finding the FDR via permutation.**

- Compute q-values of actual data
- compute q-values of permuted data (swap the order of the individuals for the genotypes when testing for associations)
- At each threshold $t(x)$ compute the approximate FDR:

$$FDR = \frac{\text{permuted q-values} \geq t(x)}{\text{actual q-values} \geq t(x)}$$

$$\approx \frac{FP}{FP + TP}$$

Returning to our problem of identifying associations for a quantitative trait, we consider the linear regression model, $y_i = x_i^T \cdot \beta$, where $i = 1, ..., n$ and $\beta = [\beta_0, \beta]$ (β_0 is the intercept term, β is the slope of the line, or the effect size of the association).

Then we can specify our hypothesis testing problem:

$H_0 : \beta_j = 0$, i.e., there is no association between the SNP and the phenotype,

$H_A : \beta_j \neq 0 \leftarrow$ two-sided, i.e., there is an association between the SNP and the phenotype.

Under the null, this sampling distribution of $\beta|x, y$ is the t-distribution with $n-2$ degrees of freedom (in this case). To compute the p-values to compare these two hypotheses, we can perform a two-sided $t - test$, which calculates the probability of seeing this estimate of β (for a specified number of degrees of freedom) under the null distribution (i.e., when there is no association between x and y).