On estimating the number of samples

1

Suppose we have two sets of samples, of size n each (control, treatment, say) and that we use a certain test to distinguish them. We'll call H_0 the null hypothesis that a given SOMAmer does not distinguish the two groups, and H_1 the alternative (think: t-test and $\mathrm{H}_0 : \mu_0 = \mu_1$) Let:

- n the number of samples in each group (as stated above)
- S be the number of SOMAmers
- T the number for which there is a real effect (H_0 does not hold.) We'll assume these are the first T.
- α level of significance that we'll use to reject each individual H_0
- $d_i, i = 1 \dots S$ the effect of the difference between the two groups, for each SOMAmer i.

Given the observations, n, α, and d_i we get the power $1 - \beta_i$ of the test, to detect an effect of size d_i in the i-th SOMAmer.

Now, since β_i is the Type II error rate, we have that the expected number of True Positive tests is given by $\mathrm{TP} = \sum_{i=1}^{T}(1 - \beta_i) = T(1 - \beta_M)$, where $\beta_M = \frac{1}{T}\sum_{i=1}^{T}\beta_i$, the average of the β for which the alternative holds. (This should be intuitively clear; but just in case: whether each of the first T tests will be positive has a probability $1 - \beta_i$, so the number of positives follow a Poisson Binomial distribution, whose expected value is given by the sum of the probabilities of each test...) Note that $1 - \beta_M$ is the average power.

On the other hand, the false positives (FP) come from the SOMAmers where H_0 holds (there are $S - T$ of them), and each one of them has a α probability of being found significant; hence we expect $\mathrm{FP} = \alpha(S - T)$

Putting these together, we can compute any kind of error we want; in particular we can get

$$\text{FDR} = \frac{\text{FP}}{\text{P}} = \frac{\text{FP}}{(\text{TP} + \text{FP})} = \frac{\alpha(S-T)}{\alpha(S-T) + (1-\beta_M)T}$$

A slightly different way of writing it (if we use π_0 for the proportion of true null hypothesis, $\pi_0 = (S-T)/S$):

$$\text{FDR} = \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta_M)(1-\pi_0)}$$

In this, the β_M depends on α, d_i and n. So if we want a sample size to control the FDR, we would need to invert this formula (tabulate it.)

2

From equation 1 we can take a different road and try to estimate the effective power $1-\beta_M$. For that, we would need estimates for π_0 and FDR, but that's exactly what Storey's algorithm does. Assuming these estimates, we get

$$1 - \beta_M = \frac{\alpha\pi_0}{1-\pi_0}\frac{1-\text{FDR}}{\text{FDR}}$$

Actually, this is kind of silly... if we set a cutoff at α we know the number of positive tests P, so if we have an estimate of π_0 we can compute this effective power directly as

$$(1-\beta_M) = \frac{\text{TP}}{\text{T}} = \frac{\text{P} - \text{FP}}{\text{T}} = \frac{\text{P} - \alpha\pi_0\text{S}}{\text{T}} = \frac{\text{P/S} - \alpha\pi_0}{(1-\pi_0)}$$

so that the only information we need is an estimate of π_0.