Studies involving multiple outcomes (e.g. genomics, MRI)

* key concept: p-values are uniformly distributed under H0
* difference between studies that are hypothesis-generating vs. hypothesis-testing
* Bonferroni always works, but at the expense of a large number of Type II errors (i.e. low power or too conservative)
* Alternative approaches:
  + Holm’s correction: Bonferroni step-down procedure
    - we compare our smallest p-value to α/n, if this is significant, we reject H0 for this test and then continue, comparing next smallest p-value to α/(n – 1)
    - continue until first fail-to-reject H0
  + Hochberg’s correction: Bonferroni step-up procedure
    - multiply highest p-value by 1, next highest by 2, etc.
    - slightly more powerful than Holm’s
    - similar to false discovery rate (FDR)
  + False Discovery Rate (Benjamini & Hochberg 1995):
    - sensitivity: P(T+/D+) vs. positive predictive value (PPV): P(D+/T+)
    - FDR is analogous to PPV: P(H0 is true / H0 was rejected)
    - Two handy features:
      1. FDR is calculated based on the p-values from the multiple tests
      2. more appropriate in cases where we expect H0 to be false for a fair number of features (e.g. genetic studies)
      3. If we choose an FDR of 0.05, it means that we would like no more than 5% of the features that we classify as significant to be false positives
         * This is not the same as choosing the FWER to be 5%
         * As with FWER, we can either use it to set a criterion or we can use it to adjust our p-values
    - FDRi = (# of false positives) / (# called significant)
      1. sort p-values and find the ith ordered p-value (= pi)
      2. if we chose this p-value as the cut-off for significance, we would have i significant p-values (by definition, due to sorting)
      3. if all H0’s are true (**uniform distribution of p-values**), then the expected # of false positives would be: n x pi
      4. Thus FDRhati = n x pi / i
      5. To control at a specific value of FDR (called q\*), we simply find the largest value of i (call this value ‘k’) for which P(i) ≤ (i/n) x q\*
         * then reject all H(i) i = 1, 2, . . . , k
      6. If all H0 are true, then FDR = FWER
      7. FDR is appropriate when a single falsely rejected H0 (i.e. false positive) is not too dangerous – see my MATLAB code: FDRdemo.m

**False Discovery Rate and the q-value:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Called significant  (H0 rejected) | Called not significant  (H0 accepted) | **Total** |
| H0 True | F  False Positive  "False Alarm"  Type I error | m0 – F  True Negative  "Correct Rejection" | **m0** |
| HA True  (H0 False) | T  True Positive  "Hit" | mA – T  False Negative  "Miss"  Type II error | **mA** |
| **Total** | **S** | **m – S** | **m** |

The table above helps us think about hypothesis testing in general. It gives all four possible outcomes of such a test, with the rows corresponding to the "ground truth" and the columns corresponding to how our test classifies the comparison. [**Note**: This same formalism is used in "Signal Detection Theory," which was developed by physicists and is now heavily used in the psychology of perception. I've included the corresponding terms in quotation marks.]

Most folks are comfortable with P-values, which give the probability of wrongly rejecting H0, also known as the *false positive rate (FPR)*. A related, but critically different, measure is called the *false discovery rate (FDR)*, which gives the proportion of tests labeled as "significant" when, in fact, H0 was true. In both cases, the numerator is the number of false positives, *F*, but, for the former, the denominator is the upper row total (m0), and, for the latter, it is the left column total (S).

**FPR** = F / (F + m0 – F) = F / m0 **FDR** = F / (F + T) = F/S

Related values that are also used to discuss this issue:

*sensitivity* = T / mA *specificity* = (m0 – F) / m0

**Q-value**:

* an extension of the original Benjamini & Hochberg approach
* rather than using the total # of tests to calculate the number of false positives, Storey estimates the number of null hypotheses using the data (Empirical Bayes!):
  + FDRhati = minp≥pi [(nhat0 x pi) / i]
  + a.k.a. q-value
  + The derivation is relatively straightforward and can be found in Storey (2002) or Storey & Tibshirani (2003)

The same table above using the annotation of Storey 2001 (classic on FDR)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Called significant  (H0 rejected) | Called not significant  (H0 accepted) | **Total** |
| H0 True | V | U | **m0** |
| HA True  (H0 False) | S | T | **m1** |
| **Total** | **R** | **W** | **m** |