Understanding q-values as a More Intuitive Alternative to p-values

Patrick Kimes

Adviser: Professor Jo Hardin

April 10, 2009

Contents

1 Background **3**

1.1 Hypothesis testing 3

1.2 Testing error 3

1.3 Multiple hypothesis testing and associated error 4

2 p-values and q-values **6**

2.1 **Cutoffs** 6

2.2 **p-values** 8

2.3 **q-values** 9

2.3.1 **Theorem 1** 10

2.3.2 **Theorem 2** 12

3 Estimation of the pFDR **18**

4 Future Work **19**

1 Background

1.1 Hypothesis testing

When an assumption concerning a parameter in a population or model is made, hypothesis tests are often used to make statistical inferences. A standard statistical hypothesis test is structured with a null and an alternative hypothesis, where the null is the assumption being tested, and the alternative contains another set of values for the parameter. For example, we may be interested in the fairness of a standard two sided coin, and thus our null hypothesis would be that the two sides of the coin come up with equal probability, and our alternative hypothesis would state that they do not. Hypothesis tests rely on some function of the observed data, called the test statistic, to determine whether the observed data are more consistent with the null or alternative hypothesis. A decision is made whether to reject or not reject the null hypothesis in favor of the alternative. In the coin example, we may flip our coin one hundred times and see the two sides in reasonably equal proportions. Having observed these data, we would determine that we should not reject our null hypothesis, as the likelihood of it occurring given that the coin is fair, is not significantly small enough.

1.2 Testing error

Hypothesis	*Fail to Reject*	*Reject*	*Total*
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

Table 1: Outcome when testing m total hypotheses

Hypothesis testing may lead to any of four possible outcomes: failing to reject when the null hypothesis is true, failing to reject when the alternative hypothesis is true, rejecting when the null hypothesis is true, and rejecting when the alternative hypothesis is true. We define the frequency at which each occurs to be: U, T, V, S, respectively. The relationship is outlined in Table 1. Both U and V occur when the null hypothesis is true and T and S when it is not. Additionally U and T are cases in which the hypothesis test leads to the null hypothesis not being rejected, and V and S when it is. In two of these cases (V & T), the statistician calculating the test is making an error.

The two types of errors hypothesis testing may lead to are Type I and Type II errors. The first occurs when the null hypothesis is incorrectly rejected, V, and the second when the null hypothesis is incorrectly not rejected, T. A Type I error, or false positive, can only occur when the null hypothesis is rejected, and similarly a Type II error, a false negative, may only occur when the null hypothesis is not rejected. Thus, in the previously discussed example of the coin where we failed to reject the null hypothesis in favor of the alternative, if the null hypothesis was actually incorrect, we have committed a Type I error. In most cases of hypothesis testing, the practitioner is more interested in constraining Type I error than Type II error, as committing a false positive is often more damaging. It can be thought of as being similar to a criminal trial, where the repercussions of incarcerating or sentencing an honest civilian are far greater than releasing a true criminal. In both hypothesis testing and criminal trials, similar methods of innocent until proven guilty are used, as the null hypothesis is not rejected unless we have sufficient evidence to do so.

1.3 Multiple hypothesis testing and associated error

In multiple hypothesis testing where several hypothesis tests are simultaneously evaluated, controlling error rates becomes substantially harder. Returning to our previous example, if we were to test ten individual two sided coins, the chances of making at least one error would be higher than if we were only testing one coin. In fact, as the number of hypotheses being tested increases, so does the overall chance of making an error. Let us illustrate this through a simple example of a multiple hypothesis test with two independent tests. Below, allow V to represent the total number of Type I errors, and each T_i be the result of the i-th test ($i = 1, 2$), with 1 being a rejection of the null, and 0 a failure to reject the null. Additionally, let α be the probabilities of making a Type I error

in any single hypothesis test, and g the probability of rejecting the null in both tests.

$$\begin{aligned}
\Pr(V \geqslant 1) &= 1 - \Pr(V = 0) \\
&= 1 - \Pr((T_1 = 0) \cap (T_2 = 0)) \\
&= 1 - (1 - \Pr((T_1 = 1) \cup (T_2 = 1))) \\
&= 1 - (1 - (\Pr(T_1 = 1) + \Pr(T_2 = 1) \\
&\quad -\Pr(T_1 = 1 \cap T_2 = 1))) \\
&= 1 - (1 - (\alpha + \alpha - g)) \\
&= \alpha + (\alpha - g) \\
&\geqslant \alpha
\end{aligned}$$

Similar calculations can be used to determine that the probability of making one Type I error with one hypothesis test is equal to just α. Therefore, the probability of incorrectly rejecting at least one null hypothesis is greater when we have two hypothesis tests than when we only have one. Multiple-testing procedures are developed specifically to address issues of testing multiple hypotheses, and provide measures of error more appropriate for multiple-testing, as well as methods for controlling and estimating them. Several methods are presented below, and one in particular, the positive false discovery rate, will be expanded upon throughout the remainder of this paper.

1. **familywise error rate (FWER)**

   The FWER is defined as the probability of making at least one Type I error while testing multiple hypotheses. It is important to note that the FWER is equal to the Type I error rate when testing only a single hypothesis. Using an example of flipping 10 two sided coins 10 times each, assuming that all of the coins are indeed fair, the chance that we will make at least one Type I error is equal to the chance that at least one coin will show behavior that is unusual enough to suspect that it is not indeed fair (e.g. landing on heads 9 out of the 10 times).

   Using Table 1, FWER is calculated by:

$$\text{FWER} = \Pr(V \geqslant 1) \tag{1}$$

2. **false discovery rate (FDR)**

   Developed more recently by Benjamini and Hochberg, the FDR is an alternate measurement of overall error that is more liberal and powerful than the FWER [1][3]. The FDR is the expected number of type I errors among all the hypothesis tests.

   Using Table 1, it is calculated by:

$$FDR = E\left(\frac{V}{R}\middle| R > 0\right) Pr(R > 0) \tag{2}$$

3. **The positive false discovery rate (pFDR)**

   Storey introduces a variation of the FDR known as the pFDR [2]. The pFDR is the calculated FDR assuming that there is at least one positive hypothesis test. Equivalently, it is the expected number of type I errors given that we have at least one positive hypothesis test.

   Using Table 1, it is calculated by:

$$pFDR = E\left(\frac{V}{R}\middle| R > 0\right) \tag{3}$$

2 p-values and q-values

2.1 Cutoffs

Error rates in hypothesis testing depend on the cutoffs used to determine the significance of the sample data set. In other words, our Type I and Type II error rates will greatly depend upon whether we determine a data set to be significant when it is only a little unusual or when it is incredibly unlikely. The extremeness of the observed data is determined as the distance from the expected results under the null hypothesis, called the test statistic, T. Cutoffs are values that are set such that all test statistics more extreme than the cutoff are considered significant if $T \in \Gamma$. The collection of all values more extreme than the cutoff comprise the rejection region for that test. Consider an example where we are flipping a two sided coin 10 times. If our null hypothesis is that the coin is fair, our expected result is 5 heads and 5 tails. If our alternative hypothesis is that the coin, on average, lands on heads more often than tails, we would want our cutoff to appropriately

reflect this. It would therefore make sense that the cutoff would be set so that it would only cause us to reject the null hypothesis in favor of the alternative when the number of heads is significantly greater than the number of tails ($T = \#\text{heads}$). Suppose we first set the cutoff to be 7 heads. Since all values further away from the expected number of heads and tails are considered significant, 7 or more heads is now the criteria for rejecting the null hypothesis. Thus, all possible combinations of tosses with 7 or more heads will comprise what is referred to as the rejection region, $\Gamma = [7, 10]$. It is not too difficult to determine, assuming the fairness of the coin, that we would be falsely rejecting the null hypothesis 176 out of 1024 times, or approximately 17% of the time. Suppose we think our cutoff results in a Type I error rate that is too large, we may change our cutoff to now only reject the null hypothesis when the number of heads is greater than or equal to 8, $\Gamma = [8, 10]$. The rejection region will also effectively change to only the set of combination of tosses with 8 or more heads. We have now reduced our Type I error rate to approximately 5% simply by altering the cutoff value and rejection region.

To calculate the Type I error rate it is necessary to assume that the null hypothesis is indeed true. The Type I error rate is the frequency of rejected null hypotheses specifically in cases when the null hypothesis is correct. If in fact we did not make this assumption on the truth of the parameter in question, the frequency would illustrate the rate of rejecting null hypothesis tests, regardless of their true nature (which would be seemingly impossible to calculate). Similarly, the proportion of significant tests assuming that the null hypothesis is actually truly false, is the power of the test. The probabilities at which Type I or Type II errors occur are examples of conditional probabilities, probabilities of events calculated with conditional assumptions. The conditional probability of an event A given an assumption, B, can be written as:

$$\Pr(\mathrm{A}\big|\mathrm{B}) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{4}$$

Conditional probabilities are important for correctly understanding and interpreting the results of hypothesis tests, as error rates can often be expressed as conditional probabilities. It is not too hard to see that the probability of a Type I error can be written as:

$$\Pr(\text{Reject Null Hypothesis}\big|\text{Null Hypothesis True}) \tag{5}$$

In the above example of 10 coin flips, the Type I error rates are calculated exactly in this way as the conditional probability:

$$\Pr(\text{At Least } n \text{ Heads} \big| \text{The Coin is Fair}) \tag{6}$$

2.2 p-values

The p-*value* is a commonly used measure for determining the significance of a hypothesis test. It is the probability of observing data as or more extreme than the data being analyzed given the null hypothesis is true. The definition of the p-value is equivalent to that of the Type I error rate of the observed data. The definition of the p-value is helpful as it offers a fairly straightforward method for interpreting the results of a test: the rate at which we would see Type I errors if we used our observed data as the significance cutoff, given the null hypothesis is true. Let X_i for $i = 1, ..., m$ be vectors of observed data for m independent tests. For all tests i, H_i is an indicator variable so that when $H_i = 0$ the null hypothesis for test i is true, and when $H_i = 1$, the alternative instead is true. The p-value can then be expressed as:

$$p\text{-value} = \Pr(\text{Data as or more extreme} \big| H = 0) \tag{7}$$

Let t_i for $i = 1, ..., m$ be the test statistics, a measure of data divergence from expected data under the null hypothesis, for the m hypothesis tests comprising the multiple hypothesis test. As stated above, the p-value for the i-th test is the probability of the data being as or more extreme than what was observed in test i, conditional on the null hypothesis being true. Equivalently, the p-value is the probability of a test statistic as more or extreme (greater) than t_i, given that the null hypothesis is true. Assuming the same notation for H_i, and using t_i as the test statistic, p-value(t_i) can be written as the conditional expression:

$$p\text{-value}(t_i) = \Pr(T \geqslant t_i \big| H_i = 0) \tag{8}$$

$\Gamma(t_i) = [t_i, \infty)$ is the rejection region, and therefore the p-value can be defined equivalently as the probability of the test statistic t_i being contained in the rejection region given that the null hypothesis is true.

$$p\text{-value}(t_i) = \Pr(T \in \Gamma(t_i) \big| H_i = 0) \tag{9}$$

2.3 q-values

Recall that the p-value is the Type I error rate when the observed data is used as the significance cutoff. The q-value for the i-th test is the pFDR when the observed data is used as the significance cutoff. For the i-th test with respective test statistic t_i, this can be written as:

$$q\text{-value}(t_i) = \text{pFDR}(T \geqslant t_i) \tag{10}$$

It is possible to show that the q-value as a function of the test statistic t_i can additionally be written:

$$q\text{-value}(t_i) = \Pr(H_i = 0 \big| T \geqslant t_i) \tag{11}$$

This theorem allows for the q-value of an observed set of data to be understood as a simple conditional probability. Expressed as the probability of the null hypothesis being true given the observed data, the q-value in eq.(11) is very similar to the p-value in eq.(8). Indeed, the q-value is the p-value with the conditional statement and event of interest reversed.

The implications of this subtle distinction can be best understood through an example. If a giant meteor where heading towards the earth a hypothesis interesting to test might be that the world was ending soon. A p-value calculation would produce the probability of an event as catastrophic or more than a meteor heading towards the earth happening give that the world was ending. The q-value would be the probability of the world ending given the observation of an event as catastrophic as a meteor heading towards Earth. In this case, the q-*value*, rather than the p-value answers a much more natural question. It makes substantially more intuitive sense to be interested in how likely it is that the world will end than how likely the event of the meteor was given the world is indeed ending. It can therefore be argued that the conditional definition of the pFDR makes it

even more intuitive than the p-value.

However, why the conditional definition of the q-value in eq.(11) follows from eq.(10) is not immediately obvious. Below is an outline of Storey's proof to show the equivalence of the two forms [2].

2.3.1 Theorem 1

Storey [2]

Suppose m identical hypothesis tests are performed with the statistics $T_1, \cdots, T_m$ and significance region Γ. Assume that (T_I, H_i) are i.i.d. random variables, $T_i \big| H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$ for some null distribution F_0 and alternative distribution F_1, and $H_i \sim Bernoulli(\pi_1)$ for $i = 1, \cdots, m$. Then

$$pFDR(\Gamma) = Pr(H = 0 \big| T \in \Gamma),$$

where $\pi_0 = 1 - \pi_1$ is the implicit prior probability used in the above posterior probability.

Proof 1 Using the definition of the pFDR from eq.(3), eq.(10) can be rewritten as the expected value of the proportion of incorrectly rejected tests, assuming at least one test is rejected, using t_i as the threshold for rejection. Let Γ_i be the corresponding test statistic rejection region. Then, it is possible to write the q-value for test i using notation from Table (1) as follows:

$$\begin{aligned} q\text{-value}(t_i) &= \text{pFDR}(\Gamma_i) && (12) \\ &= E\left[\frac{V(\Gamma_i)}{R(\Gamma_i)} \big| R(\Gamma_i) > 0\right] && (13) \end{aligned}$$

Since eq.(13) is conditional on there being at least one significant test, it can be broken up into several smaller cases of when the number of significant tests is anywhere between 1 and m. Therefore eq.(13) can be expressed as the weighted sum of the expected proportion of incorrectly rejected tests when k tests are rejected, and $k = 1, .., m$. Because all of the test statistics, t_i are independently identically distributed, it is possible to simply drop the subscript and assume the same expected value holds for all the values of i.

Each conditional expectation is weighted by the probability of it's occurrence given the original

constraint, that there is at least one significant test in the multiple testing procedure. In this case, for every possible number of significant tests, $R(\Gamma) = k$, the expected proportion of significant tests that are falsely identified is weighted by the probability of that number of tests being significant given that there is at least one significant test, $\Pr(R(\Gamma) = k \big| R(\Gamma) > 0$.

$$\begin{aligned} q\text{-value}(t) &= \sum_{k=1}^{m} E\left[\frac{V(\Gamma)}{R(\Gamma)}\Big| R(\Gamma) = k\right] \Pr(R(\Gamma) = k \big| R(\Gamma) > 0) \quad (14) \\ &= \sum_{k=1}^{m} E\left[\frac{V(\Gamma)}{k}\Big| R(\Gamma) = k\right] \Pr(R(\Gamma) = k \big| R(\Gamma) > 0) \quad (15) \end{aligned}$$

Using the basic principles of expected values, the $\frac{1}{k}$ can be removed from the expected value in eq.(15) so that the q-value can be written as:

$$q\text{-value}(t) = \sum_{k=1}^{m} \frac{1}{k} E\left[V(\Gamma)\big| R(\Gamma) = k\right] \Pr(R(\Gamma) = k \big| R(\Gamma) > 0) \quad (16)$$

Now examine the expected value, $E\left[V(\Gamma)\big| R(\Gamma) = k\right]$ in eq.(16). The expected number of Type I errors given a specified k number of significant tests, would be for example, if the test were again with 10 coins being flipped 10 times each, and $k = 3$, intuitively just the sum of the likelihoods of each of the 3 significant test being Type I errors. Because $V(\Gamma)$ is the number of Type I errors, it is the number of tests that are both following the null distribution and contained in the test statistic rejection region. Let $\mathbb{1}(x)$ be an indicator function that returns 1 when the argument x is true, and 0 when it is not. Using these indicator functions, it is possible to simply ask of each test i whether both are equal to 1, and if so then test i is a Type I error. If the samples are arranged such that samples X_i, $i = 1, ..., k$ have test statistics t_i such that they are significant, and samples X_i, $i = k+1, ..., m$ have test statistics t_i such that they are not significant, the expected value can be rewritten:

$$E\left[V(\Gamma)\big| R(\Gamma) = k\right] = E\left[\sum_{i=0}^{m} \mathbb{1}(t_i \in \Gamma)\mathbb{1}(H_i = 0 \Big| T_1, \cdots, T_k \in \Gamma; T_{k+1}, \cdots, T_m \notin \Gamma\right] \quad (17)$$

Since $\mathbb{1}(t_i \in \Gamma)$ is 1 for only $i = 1, ..., k$, all other values of i can be dropped from eq.(17).

$$E\left[V(\Gamma)\middle|R(\Gamma) = k\right] = E\left[\sum_{i=0}^{m} \mathbb{1}(H_i = 0)\middle|T_1, \cdots, T_k \in \Gamma; T_{k+1}, \cdots, T_m \notin \Gamma\right] \tag{18}$$

Using again simple rules of expected values, the expectation can be brought into the summation so that:

$$E\left[V(\Gamma)\middle|R(\Gamma) = k\right] = \sum_{i=0}^{m} E\left[\mathbb{1}(H_i = 0)\middle|T_i \in \Gamma\right] \tag{19}$$

The expected value of the indicator function in eq.(19) given that the test statistics falls within the rejection region is just the probability of the null hypothesis being true given the same constraint. Since the test statistic of each data set is independently and identically distributed, for the k such test statistics, the expected values are all the same. The sum of the k expected values is:

$$E\left[V(\Gamma)\middle|R(\Gamma) = k\right] = k \cdot \Pr(H = 0\middle|T \in \Gamma) \tag{20}$$

By taking eq.(20) and substituting it in to eq.(16), the following conclusion can be reached:

$$\begin{aligned} q\text{-value}(t) &= \sum_{k=1}^{m} \frac{k \cdot \Pr(H = 0\middle|T \in \Gamma)}{k} \Pr(R(\Gamma) = k\middle|R(\Gamma) > 0) && (21) \\ &= \Pr(H = 0\middle|T \in \Gamma) && (22) \end{aligned}$$

The q-value was first introduced as an extension of the pFDR in an attempt to find an error measure similar to the p-value for the pFDR. The pFDR is a measure of false discovery rates, and t the q-value for a hypothesis test is the pFDR when using the observed data as the significance cutoff. It is not appropriate to use p-values in the context of the pFDR because of the conditional definition of the pFDR.

2.3.2 Theorem 2

Storey[2]

For m identical hypothesis tests, $pFDR^T(\Gamma_\alpha) = pFDR^P(\{p : p \leqslant \alpha\})$, which implies that the q-

value can be calculated from either the original statistics or their p-values. Also, when the statistics are independent and follow a mixture distribution then

$$q\text{-}value(t) = pFDR^P(\{p : p \leqslant p\text{-}value(t)\}) \tag{23}$$

if and only if $G_1(\alpha)/\alpha$ *is decreasing in* α*.*

Proof 2 Assume there are m independent and identical hypothesis tests. Let $0 < \alpha < 1$ be the p-value significance cutoff and Γ_α denote the corresponding significance region containing all test statistics with p-values less than or equal to α. Suppose test statistic t has a p-value less than α. By definition of Γ_α, $t \in \Gamma_\alpha$. Conversly, suppose there is a test statistic $t \in \Gamma_\alpha$. The corresponding p-value will also be less than α. Assume that the test statistics t_i, $i = 1, ..., m$ are independent and identically follow a mixture distribution, meaning that at some prior probability, π_0, an individual data set is truly distributed according to the null hypothesis and at prior probability, $\pi_1 = (1 - \pi_0)$, distributed according to the alternative hypothesis. What will follow is a proof that the rejection region used to calculate the q-value for a test i from the test statistics of the data sets corresponds to the rejection region for the q-value calculated based on the p-values of the same data sets. This is interesting as common knowledge about the q-value, pFDR and how they are calculated hardly exists in contrast to similar information about p-values.

Let G_1 and G_0 be functions of the level α defined respectively as:

$$G_1(\alpha) = \int_{\Gamma_\alpha} dF_1 = \Pr(T \in \Gamma_\alpha \big| H = 1) \tag{24}$$

$$G_0(\alpha) = \int_{\Gamma_\alpha} dF_0 = \Pr(T \in \Gamma_\alpha \big| H = 0) \tag{25}$$

Since $G_1(\alpha)$ is the probability of correctly rejecting the null hypothesis when the alternative is true, it is the power of the test at significance level α. $G_0(\alpha)$ is simply the probability of incorrectly rejecting the null hypothesis when it is indeed true. This is the size of the test, and therefore equal to α. The pFDR for a given α level can be rewritten as the probability of the null hypothesis being true given the test statistic is contained in the rejection region according to Theorem 1 [2].

$$\underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \ \mathrm{pFDR}(\Gamma_\alpha) = \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \ \Pr(H = 0 \big| T \in \Gamma_\alpha) \tag{26}$$

Recall that π_0 and π_1 are the respective prior probabilities of the null and the alternative hypotheses being true. By Bayes Theorem, eq.(26) is equivalent to the following:

$$\underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \Pr(H = 0 \big| T \in \Gamma_\alpha) = \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \frac{\Pr(T \in \Gamma_\alpha \big| H = 0)\Pr(H = 0)}{\Pr(T \in \Gamma_\alpha)} \tag{27}$$

The denominator $\Pr(T \in \Gamma_\alpha$, is the probability of rejecting the null hypothesis. The null hypothesis can be rejected when, either: the null hypothesis is indeed true, or when the alternative hypothesis is instead true. Additionally, the probability of any two events concurrently occurring can is equal to the probability of one the events occurring multiplied by the conditional probability of the second event occurring given the occurrence of the first is already known. Therefore:

$$\begin{aligned} \Pr(T \in \Gamma_\alpha) &= \Pr(T \in \Gamma_\alpha \cap H = 0) + \Pr(T \in \Gamma_\alpha \cap H = 1) \\ &= \Pr(T \in \Gamma_\alpha \big| H = 0)\Pr(H = 0) + \Pr(T \in \Gamma_\alpha \big| H = 1)\Pr(H = 1). \end{aligned} \tag{28}$$

Recall, π_0 and π_1 are the respective prior probabilities of the null and alternative hypotheses being true, and therefore can be used to replace $\Pr(H = 0)$ and $\Pr(H = 1)$ in eq.(28). Note that $G_0(\alpha) = \alpha$ and $G_1(\alpha)$ can be used to replace $\Pr(T \in \Gamma_\alpha \big| H = 0)$ and $\Pr(T \in \Gamma_\alpha \big| H = 1)$, respectively. By substituting eq.(28) into the denominator of eq.(27) and using the mentioned replacements, the following becomes evident.

$$\begin{aligned} \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \Pr(H = 0 \big| T \in \Gamma_\alpha) &= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \frac{\Pr(T \in \Gamma_\alpha \big| H = 0)\Pr(H = 0)}{\Pr(T \in \Gamma_\alpha \big| H = 0)\Pr(H = 0) + \Pr(T \in \Gamma_\alpha \big| H = 1)\Pr(H = 1)} \\ &= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \frac{\alpha \cdot \pi_0}{\alpha \cdot \pi_0 + G_1(\alpha) \cdot \pi_1} \qquad (29) \\ &= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \frac{\alpha \cdot \pi_0}{\alpha \cdot \pi_0 + G_1(\alpha) \cdot \pi_1} \cdot \frac{\frac{1}{\alpha}}{\frac{1}{\alpha}} \qquad (30) \\ &= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \frac{\pi_0}{\pi_0 + \frac{G_1(\alpha)}{\alpha} \cdot \pi_1} \qquad (31) \end{aligned}$$

When trying to minimize the fraction in eq.(31), π_0 and π_1 can be ignored because they are constants. What is left is the $\frac{G_1(\alpha)}{\alpha}$ in the denominator. Therefore, by maximizing the fraction in the denominator, the entire equation is effectively minimized. From this, the conclusion can be reached that the rejection region Γ_α containing the test statistic T minimizing $\frac{\alpha}{G_1(\alpha)}$ will also

minimize the pFDR.

$$\underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \; pFDR(\Gamma_\alpha) = \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \; \frac{\alpha}{G_1(\alpha)} \tag{32}$$

$$= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \; \frac{\Pr(T \in \Gamma_\alpha \big| H = 0)}{\Pr(T \in \Gamma_\alpha \big| H = 1)} \tag{33}$$

The fraction in eq.(32) can be minimized by taking the derivative of the ratio with respect to α and solving for when it is equal to 0.

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\alpha} \frac{\alpha}{G_1 \alpha)} \\ &= \frac{G_1(\alpha)(1) - \alpha({G_1}'(\alpha))}{G_1(\alpha)^2} \\ &= G_1(\alpha) - \alpha({G_1}'(\alpha)) \\ \alpha({G_1}'(\alpha)) &= G_1(\alpha) \\ \alpha &= \frac{G_1(\alpha)}{{G_1}'(\alpha)} \end{aligned}$$

$\frac{\alpha}{G_1(\alpha)}$ is minimized when $\alpha = \frac{G_1(\alpha)}{{G_1}'(\alpha)}$ and therefore so is $pFDR(\Gamma_\alpha)$.

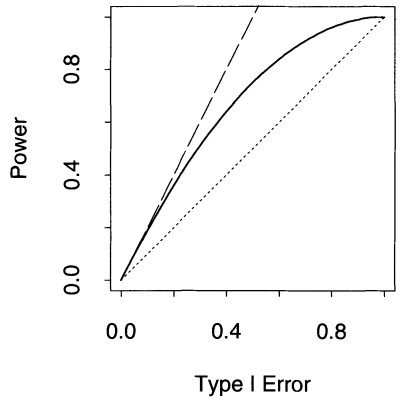

Figure 1: Graph of α versus $G_1(\alpha)$

Figure 1 shows α versus $G_1(\alpha)$ = power, for a set of significance regions including the identity function represented by a dotted line, and a line tangent to the power function and passing through the origin draw with a dashed line. The power function is concave and therefore only has one

line tangent to the curve passing through the origin. This singular tangent line is of the form: $f(\alpha) = b \cdot \alpha + 0$ because it passes through the origin, and further if α^* is the point of tangency along the power function, then the slopes of the curve and the line will be equal at the point of tangency, ${G_1}'(\alpha^*) = f'(\alpha^*)$. The slope of the line tangent to the curve at $(\alpha, G_1(\alpha))$, $f'(\alpha)$ is equal to $\frac{G_1(\alpha)-0}{\alpha-0}$. Thus:

$$\begin{aligned} {G_1}'(\alpha^*) &= f'(\alpha^*) && (34)\\ &= \frac{G_1(\alpha^*)-0}{\alpha^*-0} && (35)\\ \alpha^* &= \frac{G_1(\alpha^*)}{{G_1}'(\alpha^*)} && (36) \end{aligned}$$

(36) shows that α^* is therefore a critical point for $\frac{\alpha}{G_1(\alpha)}$ as is every other point of tangency passing through the origin. Since this relationship is known, it is easy to see that by simply finding the α^* such that the slope of the tangency, ${G_1}'(\alpha^*) = \frac{G_1(\alpha^*)}{\alpha^*}$ is maximized will at the same time also find the level at which $\frac{\alpha}{G_1(\alpha^*)}$ is minimized. Therefore, the α^* at which the slope of the line tangent to the power function is greatest is also the point at which the pFDR is minimized. Additionally, if the power function is concave, implying that the derivative of the function ${G_1}'(\alpha^*) = \frac{G_1(\alpha^*)}{\alpha^*}$ is monotonically decreasing in α, then the $pFDR(\Gamma_\alpha)$ will inversely be increasing in α. Therefore, the significance region, Γ_α containing t and minimizing $pFDR(\Gamma_\alpha$ can be found by finding the minimum α that still allows for t to be contained in the rejection region. Coincidentally, the significance region specified by the minimum α that still contains t also minimizes, α, and therefore the p-value.

$$\begin{aligned} \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \Pr(H=0 \big| T \in \Gamma_\alpha) &= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} pFDR(\Gamma_\alpha)\\ &= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \frac{\alpha}{G_1(\alpha)}\\ &= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \alpha\\ &= \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \Pr(T \in \Gamma_\alpha \big| H=0) \end{aligned} \quad (37)$$

When the power function is concave, the q-value thus shares the same test statistic significance region with the p-value. To help distinguish between the two better, T and P superscripts are used on the $pFDR$ specifying either the $pFDR$ calculated using test statistics or p-values. It is also simple to show that calculating the q-value from either the original test statistics or the p-values

will produce the same result. Because the rejection region is defined as the set of all test statistics less than the cutoff, a test statistic will only have a p-value less than or equal to the cutoff if and only if the test statistic is contained in the rejection region. Therefore,

$$pFDR^T(\Gamma_\alpha) = pFDR^P(\{p : p \leqslant \alpha\}) \tag{38}$$

Given the test statistics are independent and follow a mixture distribution,

$$q\text{-value}(t) = pFDR^P(\{p : p \leqslant p\text{-value}(t)\}) \tag{39}$$

if and only if $\frac{G_1(\alpha)}{\alpha}$ is decreasing in α. This can easily be shown by using the conclusions drawn above.. First, assume that $\frac{G_1(\alpha)}{\alpha}$ is decreasing in α, that the power function is concave, and let $\Gamma_{\alpha^*} = \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\operatorname{argmin}} \; pFDR^T(\Gamma_\alpha)$ for all test statistics $T = t$. Γ_{α^*} is thus the rejection region still containing the statistic t and minimizing $pFDR^T$. This also means that, $pFDR^T(\Gamma_{\alpha^*})$ is the q-value for the test statistic t. Additionally, by applying eq.(38),

$$\begin{aligned} q\text{-value}(t) &= pFDR^T(\Gamma_{\alpha^*}) & (40)\\ &= pFDR^P(\{p : p \leqslant \alpha^*\}) & (41) \end{aligned}$$

In eq.(37) it was shown that the rejection region minimizing $pFDR(\Gamma_{\alpha^*})$ also minimizes $\Pr(T \in \Gamma_{\alpha^*}) = \alpha^*$. Thus, the p-value is equal to the cutoff, α^* and the following is true:

$$q\text{-value}(t) = pFDR^P(\{p : p \leqslant p\text{-value}(t)\}) \tag{42}$$

Now assume that $q\text{-value}(t) = pFDR^P(\{p : p \leqslant p\text{-value}(t)\})$ is true for all t. As p-value (t) increases so will the the number of Type I errors, and therefore by the definition of the q-value, so will q-value (t). If $q\text{-value}(t) = \frac{\alpha}{G_1(\alpha)}$ is increasing in $p\text{-value}(t) = \alpha$, then it is clear that the inverse, $\frac{G_1(\alpha)}{\alpha}$ will be decreasing in α.

3 Estimation of the pFDR

Recall that Storey defines the positive false discovery rate (pFDR) for m identical hypothesis tests with independent statistics $T_1, \ldots, T_m$ and rejection region Γ as:

$$\begin{aligned} pFDR(\Gamma) &= \frac{\pi_0 Pr(T \in \Gamma \big| H = 0)}{Pr(T \in \Gamma)} \\ &= Pr(H = 0 \big| T \in \Gamma), \end{aligned} \tag{43}$$

where π_0 is the prior probability that the null hypothesis is true for any of the m identical hypothesis tests.

Storey's method of estimating the pFDR relies on rejecting based on the independent p-values. Therefore, it is necessary to redefine the pFDR as a function of p-values, where $[0, \gamma]$ is the p-value rejection region that can be simply referred to by the cutoff level, γ, $0 < \gamma < 1$. This does not change anything more than notation as for all tests with p-values less than γ, the corresponding test statistic t, $t \in \Gamma_\gamma$.

$$p\text{-value}(t) \in [0, \gamma] \textit{ iff. } t \in \Gamma_\gamma \tag{44}$$

Thus the pFDR can be rewritten as:

$$pFDR(\gamma) = \frac{\pi_0 Pr(P \leqslant \gamma \big| H = 0)}{Pr(P \leqslant \gamma)} \tag{45}$$

Because the probability of the p-value being less than the significance cut-off under the null hypothesis is by definition equal to the cut-off, $\Pr(P \leqslant \gamma \big| H = 0)$ can be replaced with γ in eq.(45).

$$pFDR\gamma = \frac{\pi_0 \gamma}{Pr(P \leqslant \gamma)} \tag{46}$$

The pFDR can therefore be estimated for any given γ by calculating estimates for both π_0 and $Pr(P \leqslant \gamma)$. where λ is some well chosen cutoff such that $\lambda \in [0, 1]$, and P is the random p-value result of any given test.

$$\widehat{\pi_0} = \frac{\#\{p_i > \lambda\}}{(1 - \lambda)m} = \frac{W(\lambda)}{(1 - \lambda)m} \tag{47}$$

The distributions of p-values for samples coming from the null distribution of data, F_0, and for those coming from the alternative distribution of data, F_1, is different. By the definition of the p-value, the p-values for data from F_0 will be uniformly distributed between 0 and 1. However, the p-values for the data from F_1 will be heavily skewed right with the majority being closer to 0. These two basic concepts are why rejecting the null hypothesis based on p-values is appropriate. The intuition behind eq.(47), therefore is that because the p-values from alternative data are so incredibly skewed right, that above a certain threshold, most if not all the tests with p-values that high will be for data drawn from the null distribution. Thus by taking the number p-values above the threshold and dividing them by the expected number of p-values above the threshold given all the tests were drawn from the null distribution gives an estimate of the proportion of tests where $H = 0$. λ is the well chosen cut-off above which it is expected that all the p-values come from data drawn the null distribution.

$$\widehat{Pr}(P \leqslant \gamma) = \frac{\#\{p_i \leqslant \gamma\}}{m} = \frac{R(\gamma)}{m} \tag{48}$$

The probability that any given hypothesis test will return a p-value contained in the rejection region, is estimated by the proportion of positive hypotheses.

$$\widehat{pFDR}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{Pr}(P \leqslant \gamma)\{1-(1-\gamma)^m\}} = \frac{W(\lambda)\gamma}{(1-\lambda)\{R(\gamma) \cup 1\}\{1-(1-\lambda)^m\}} \tag{49}$$

The estimate of pFDR given by eq.(49) simply combines the estimates of π_0 and $Pr(P \leqslant \gamma)$ and adds an additional element, $\frac{1}{\{1-(1-\gamma)^m\}}$. The additional element is included to account for the pFDR conditioning on the number of positive tests being at least one. $\frac{1}{\{1-(1-\gamma)^m\}}$ is a lower bound for $Pr(R(\gamma) > 0)$. Storey thus uses data to estimate the pFDR.

4 Future Work

Storey has further extended his work on the pFDR and q-value by producing a function in the statistical package R, qval that attempts to produce q-values based solely on data inputted p-values. The function incorporates many assumptions that have been discussed above. A larger discussion and examination of the validity of the function should be explored in the future. Through controlled simulations, how well the function performs in calculating the q-value should be tested. Further,

Storey uses a cubic spline method to estimate the λ to be used in eq.(47). A deeper examination of this choice of λ estimation can also be tested in future work [4].

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **B**, **57**, 289-300.

[2] Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* **31**, 2013-2035.

[3] Storey, J. D. (2001). A direct approach to false discovery rates. *Journal of the Royal Statistical Socety* **64**, 479-498.

[4] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Science* **100**, 9440-9445.