Estimating π_0 for High Dimensional Discrete Data and Adaptive FDR Estimation

Naomi Altman[1] & Isaac Dialsingh[2]

Joint Statistical Meetings San Diego
July 29, 2012

1. The Pennsylvania State University
naomi@stat.psu.edu

2. University of the West Indies
Isaac.Dialsingh@sta.uwi.edu

False Discovery Rate

Controlling error rates is essential for high dimensional “omics” data.

False Discovery Rate

Controlling error rates is essential for high dimensional “omics” data.

Benjamini & Hochberg (1995) realized that when testing 1000’s of hypotheses a few errors could be tolerated.

False Discovery Rate (FDR) is the expected percentage of null hypotheses among the statistically significant tests.

PENNSTATE

False Discovery Rate

Controlling error rates is essential for high dimensional "omics" data.

Benjamini & Hochberg (1995) realized that when testing 1000's of hypotheses a few errors could be tolerated.

False Discovery Rate (FDR) is the expected percentage of null hypotheses among the statistically significant tests.

Table: Outcomes of m tests.

	Not Significant	Significant	Total
True Null	U	V	m_0
False Null	T	S	m_1
Total	W	R	m

$FDR = E\left(\frac{V}{R}\right)$

R: number of rejections
V: number of false rejections

Estimating FDR

Table: Outcomes of m tests.

	Not Significant	Significant	Total
True Null	U	V	m_0
False Null	T	S	m_1
Total	W	R	m

Estimating FDR

Table: Outcomes of m tests.

	Not Significant	Significant	Total
True Null	U	V	m_0
False Null	T	S	m_1
Total	W	R	m

$$\pi_0 = \frac{m_0}{m}$$

m: number of tests
m_0: number of null tests

PENNSTATE

Estimating FDR

Table: Outcomes of m tests.

	Not Significant	Significant	Total
True Null	U	V	m_0
False Null	T	S	m_1
Total	W	R	m

$$\pi_0 = \frac{m_0}{m}$$

m: number of tests
m_0: number of null tests

Estimating π_0 is an essential part of estimating FDR.
A number of methods are available for continuous test statistics.

PENNSTATE

Estimating FDR

For discrete test statistics we want to:

- Estimate π_0.
- Estimate FDR.

PENNSTATE

Estimating FDR

For discrete test statistics we want to:

- Estimate π_0.
- Estimate FDR.

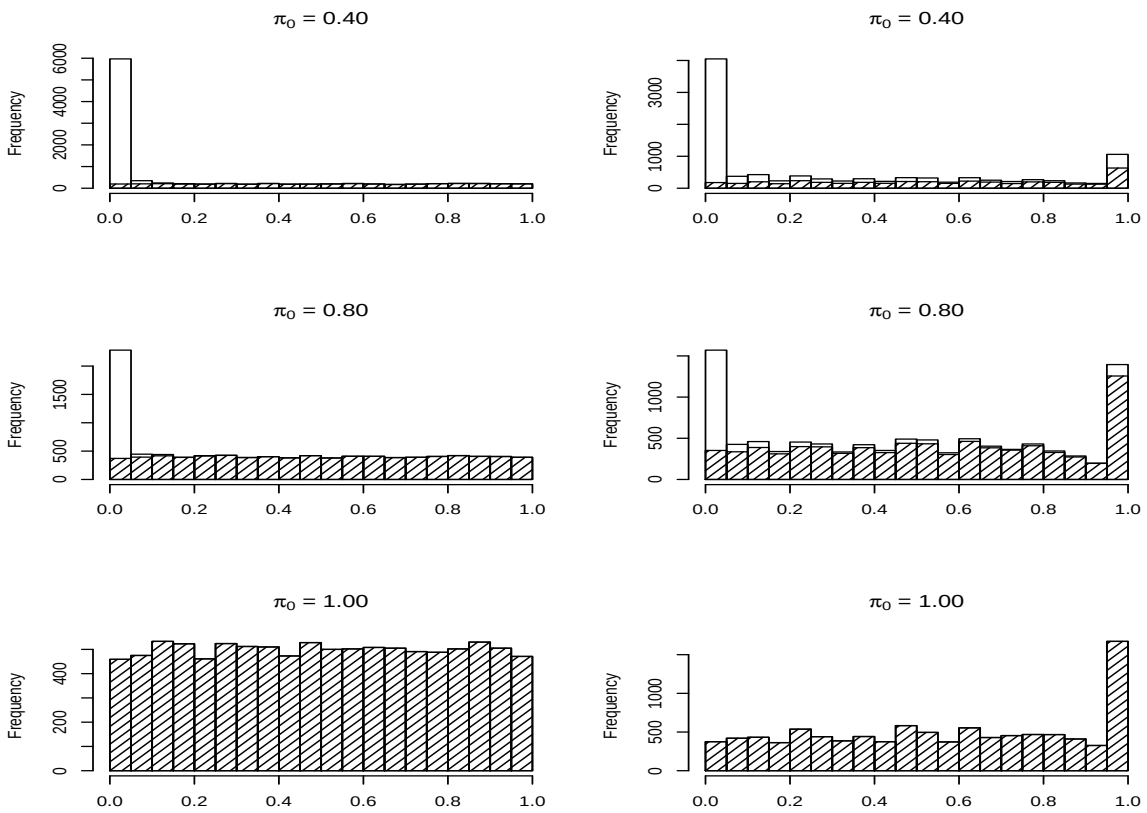
Discrete test statistics

arise from binary and count data such as

- read counts in RNA-seq and ChIP-seq
- SNP studies
- thresholding (above/below)

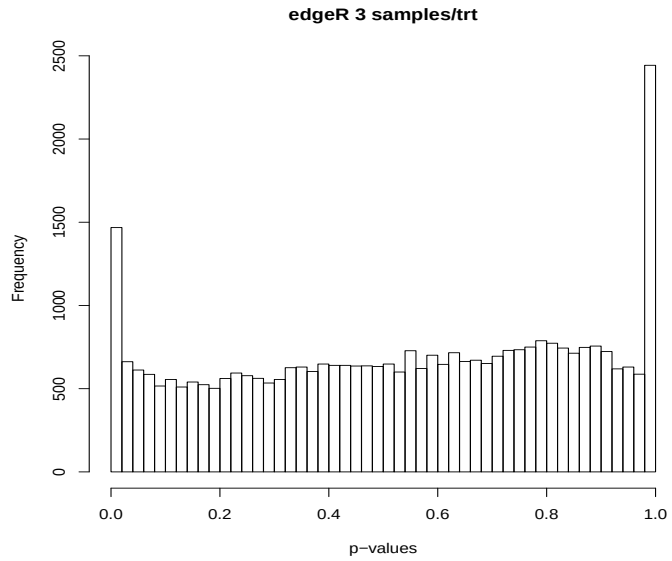
PENNSTATE

Why does discreteness matter?

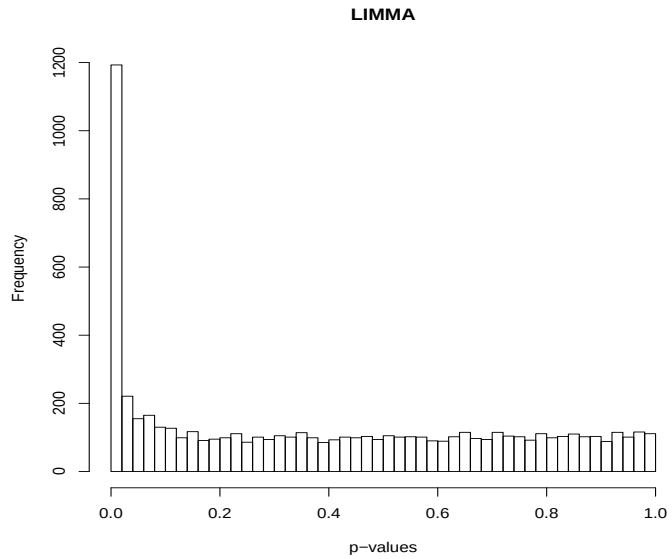


The histograms on the left are p-values from 1000 2-sample t-tests.
The histograms on the right are p-values from 1000 Fisher exact tests.
Each row has the same value of π_0.

Why does discreteness matter?



P-values from an RNA-seq study of 2 maize genotypes with 3 biological replicates.

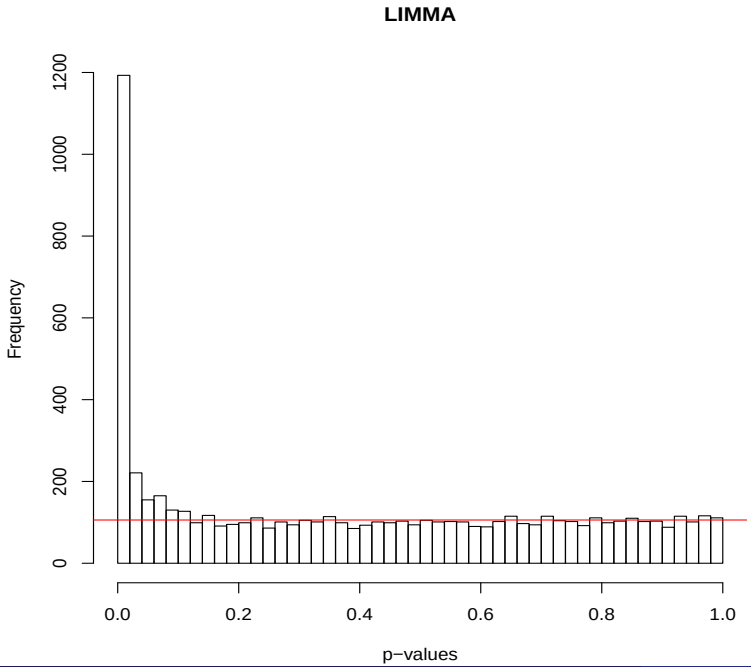


P-values from a microarray study in poppy tissues with 4 biological replicates.

Estimating π_0

Estimating π_0 from continuous p-values

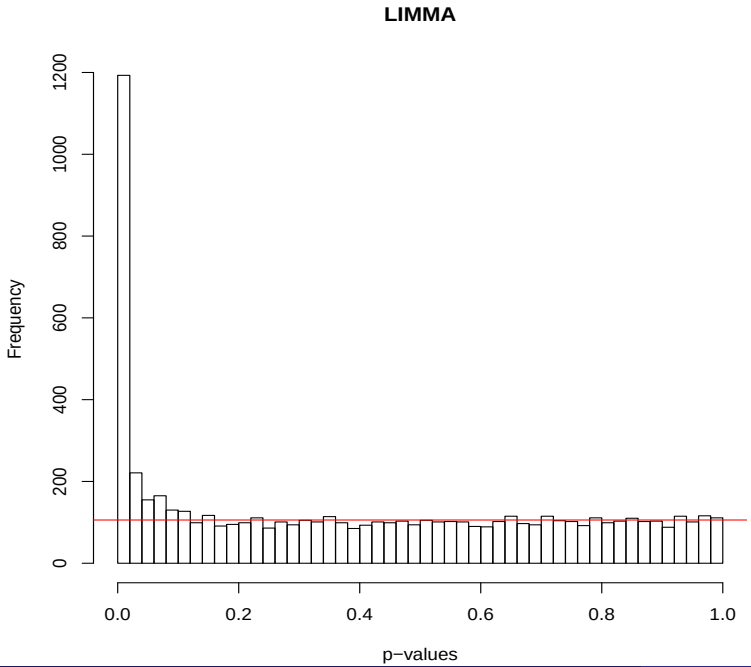
- Storey (2002) estimates height of flat part of histogram.
- Nettleton et al (2006) estimate the heights of the bins in excess of expected given $\hat{\pi}_0$.
- Pounds and Cheng (2004) assume all true non-nulls have p=0, so $2 * \bar{p} \approx \pi_0$.



Estimating π_0

Estimating π_0 from continuous p-values

- Storey (2002) estimates height of flat part of histogram.
- Nettleton et al (2006) estimate the heights of the bins in excess of expected given $\hat{\pi}_0$.
- Pounds and Cheng (2004) assume all true non-nulls have p=0, so $2 * \bar{p} \approx \pi_0$.



Estimating π_0 from discrete p-values

- These methods seem less plausible since low power non-null tests may have p-values far from 0.
- We add 3 new methods.

Estimating π_0

Estimating π_0 from discrete p-values

- There is often an ancillary statistic which determines the distribution of the test statistic. e.g. row totals.

PENNSTATE

Estimating π_0

Estimating π_0 from discrete p-values

- There is often an ancillary statistic which determines the distribution of the test statistic. e.g. row totals.
- If the ancillary statistic is known for each test, the distribution of p-values under the null is known.

PENNSTATE

Estimating π_0

Estimating π_0 from discrete p-values

- There is often an ancillary statistic which determines the distribution of the test statistic. e.g. row totals.
- If the ancillary statistic is known for each test, the distribution of p-values under the null is known.
- If there are many tests with the same value of the ancillary, the empirical distribution of the p-values can be estimated by the observed frequencies.

PENNSTATE

Estimating π_0

Estimating π_0 from discrete p-values

- There is often an ancillary statistic which determines the distribution of the test statistic. e.g. row totals.
- If the ancillary statistic is known for each test, the distribution of p-values under the null is known.
- If there are many tests with the same value of the ancillary, the empirical distribution of the p-values can be estimated by the observed frequencies.
- We may prefer to think of the ancillary statistics as random. Since we have a large number of tests, we can use the empirical distribution of the ancillary to estimate the ancillary distribution.

PENNSTATE

Estimating π_0

Estimating π_0 from discrete p-values

- There is often an ancillary statistic which determines the distribution of the test statistic. e.g. row totals.
- If the ancillary statistic is known for each test, the distribution of p-values under the null is known.
- If there are many tests with the same value of the ancillary, the empirical distribution of the p-values can be estimated by the observed frequencies.
- We may prefer to think of the ancillary statistics as random. Since we have a large number of tests, we can use the empirical distribution of the ancillary to estimate the ancillary distribution.
- We assume that π_0 is the same for all values of the ancillary.

PENNSTATE

Estimating π_0

Mixture distribution of p-values

We use the mixture distribution

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_A(p)$$

where

- f is the distribution of the p-values,
- f_0 is the distribution of p-values for the hypotheses that are truly null and
- f_A is the distribution of p-values for the hypotheses that are truly not null.

PENNSTATE

Estimating π_0 using $f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_A(p)$

Regression Method

- Regression method - regress empirical frequencies of p-values against theoretical.
- The slope is approximately π_0.

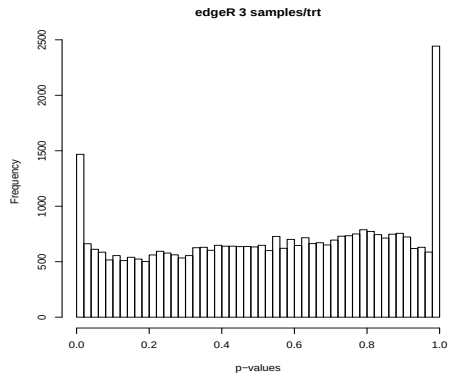
Estimating π_0 using $f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_A(p)$

Regression Method

- Regression method - regress empirical frequencies of p-values against theoretical.
- The slope is approximately π_0.
- Requires many tests with the same ancillary statistic.

PENNSTATE

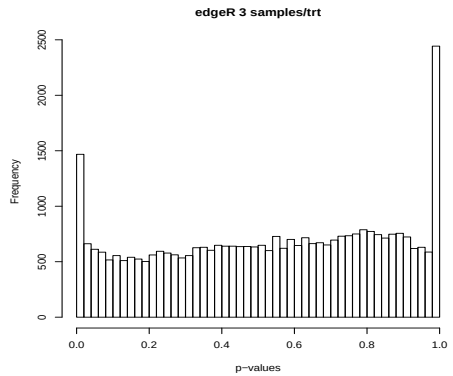
Estimating π_0 using the histogram of p-values



Methods based on the histogram of p-values

- We compare the observed histogram with the histogram expected under the null.

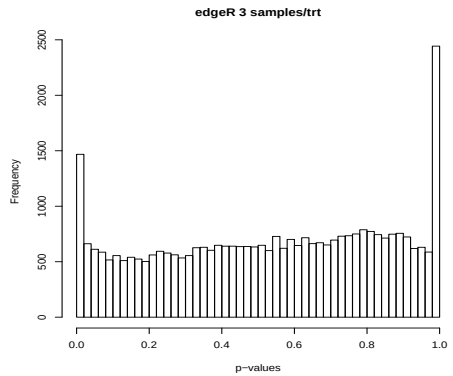
Estimating π_0 using the histogram of p-values



Methods based on the histogram of p-values

- We compare the observed histogram with the histogram expected under the null.
- For the null distribution we may
  - condition on the observed values of the ancillary
  - simulate from the distribution of ancillaries

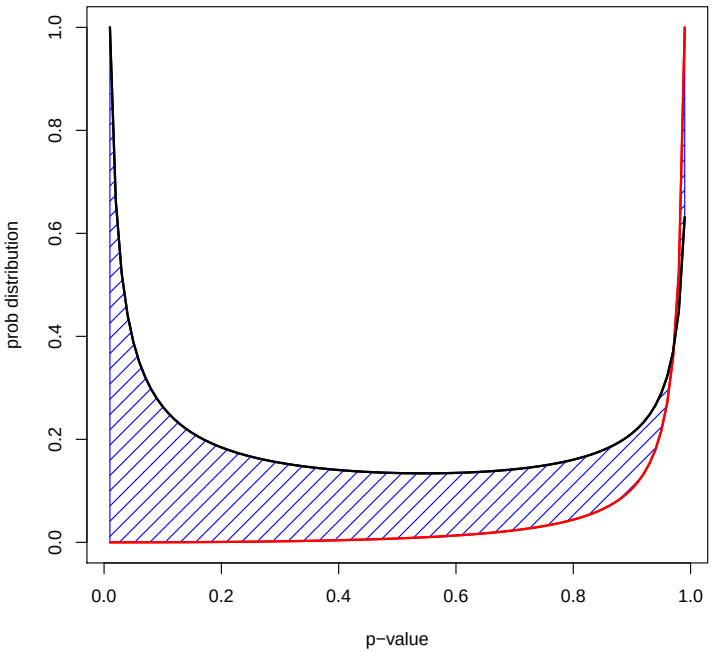
Estimating π_0 using the histogram of p-values



Methods based on the histogram of p-values

- We compare the observed histogram with the histogram expected under the null.
- For the null distribution we may
  - condition on the observed values of the ancillary
  - simulate from the distribution of ancillaries
- We want to capture peaks near 0 and 1, so we do not want equally spaced bins.
- Instead, we use equal frequency bins of $0.5f(p) + 0.5f_0(p)$.

Estimating π_0 using the histogram of p-values



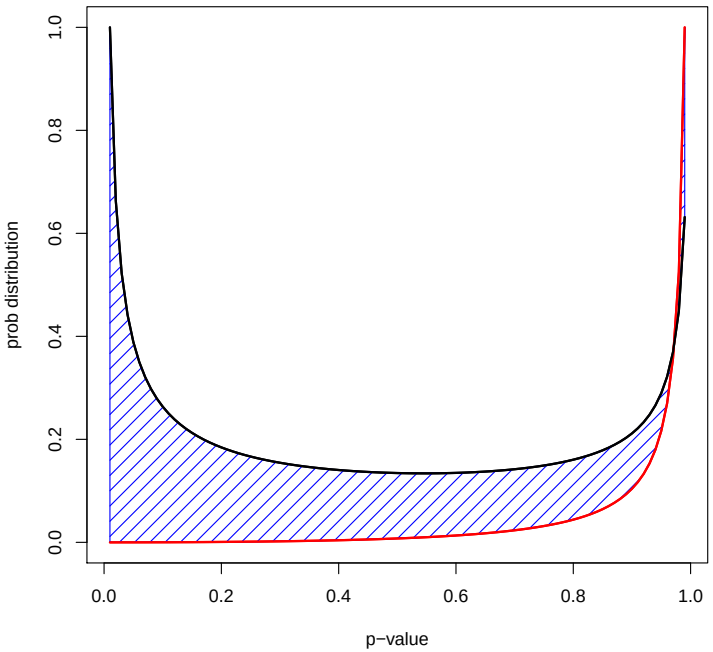
Histogram Method

- Note that

$$\begin{aligned} \int_p & |f(p) - f_0(p)|dp \\ = & (1-\pi_0)\int_p |f_A(p) - f_0(p)|dp \\ \leq & \ 2(1-\pi_0) \end{aligned}$$

so $\pi_0 \geq 1 - \frac{1}{2}\int_p |f(p) - f_0(p)|dp$.

Estimating π_0 using the histogram of p-values



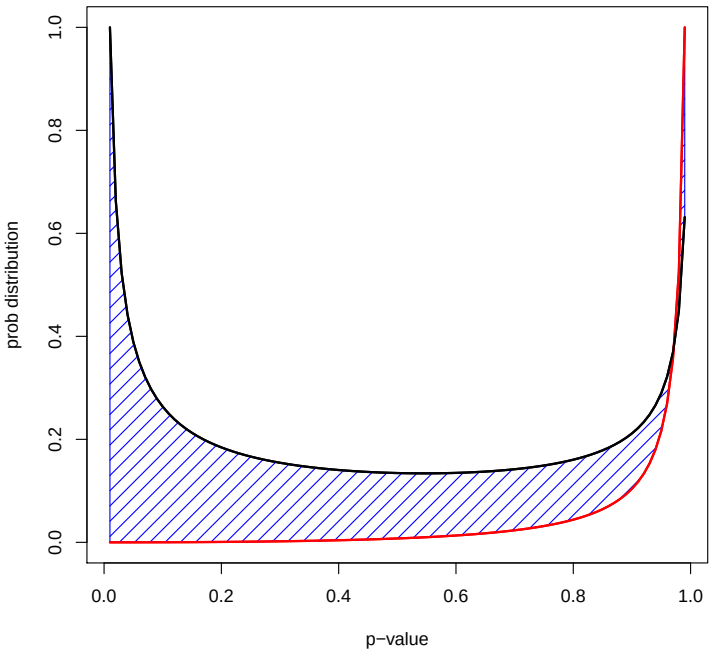
Histogram Method

- Note that

$$\begin{aligned} \int_p & |f(p) - f_0(p)|dp \\ = & (1-\pi_0)\int_p |f_A(p) - f_0(p)|dp \\ \leq & \ 2(1-\pi_0) \end{aligned}$$

  so $\pi_0 \geq 1 - \frac{1}{2}\int_p |f(p) - f_0(p)|dp$.
- Compute the bin boundaries.
- Bin both the observed and expected histograms.

Estimating π_0 using the histogram of p-values



Histogram Method

- Note that

$$\begin{aligned} \int_p & |f(p) - f_0(p)|dp \\ = & (1-\pi_0)\int_p |f_A(p) - f_0(p)|dp \\ \leq & \ 2(1-\pi_0) \end{aligned}$$

  so $\pi_0 \geq 1 - \frac{1}{2}\int_p |f(p) - f_0(p)|dp$.
- Compute the bin boundaries.
- Bin both the observed and expected histograms.
- S=the sum of the absolute difference in bin frequencies.
- $\hat{\pi}_0 = 1 - \frac{1}{2}S$

Estimating π_0 using the histogram of p-values

Hybrid Method using $f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_A(p)$

- Note that the expected bin frequencies B for bin b in the p-value histogram is

$$E(B) = \pi_0 E_0(B) + (1 - \pi_0) E_A(B)$$

PENNSTATE

Estimating π_0 using the histogram of p-values

Hybrid Method using $f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_A(p)$

- Note that the expected bin frequencies B for bin b in the p-value histogram is

$$E(B) = \pi_0 E_0(B) + (1 - \pi_0) E_A(B)$$

- Compute the bin boundaries.
- Bin both the observed and expected histograms.

PENNSTATE

Estimating π_0 using the histogram of p-values

Hybrid Method using $f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_A(p)$

- Note that the expected bin frequencies B for bin b in the p-value histogram is

$$E(B) = \pi_0 E_0(B) + (1 - \pi_0) E_A(B)$$

- Compute the bin boundaries.
- Bin both the observed and expected histograms.
- Regress the empirical frequencies of the observed histogram on against the expected frequencies under the null.
- The slope is approximately π_0.

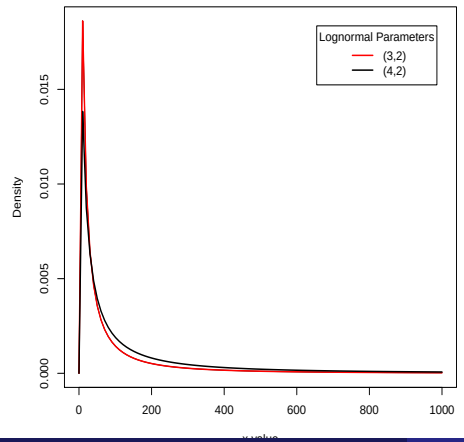
PENNSTATE

Simulated RNA-seq data

Data

We simulated RNA-seq data assuming:

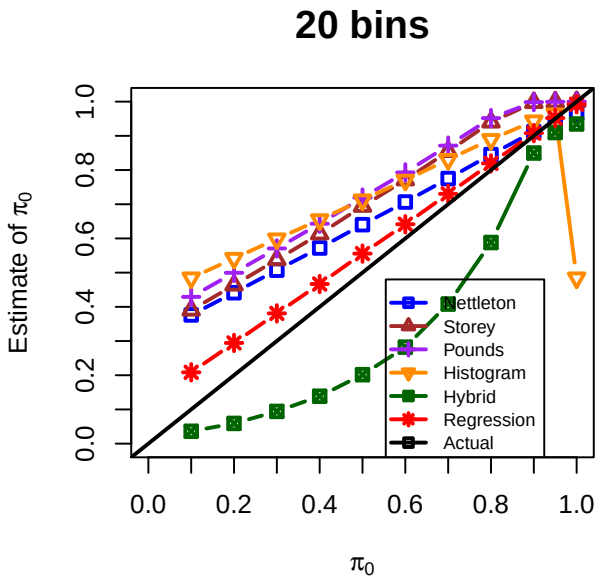
- m (number of tests) =1000 or 10,000.
- $\pi_0 = 0.1, 0.2 \cdots 0.8, 0.9, 0.95, 1.0$
- Two different discretized log-Normal distributions for total reads/feature estimated from real data.
- Features are independent within sample.
- We used 2 treatments with no replication.
- The statistic was Fisher's Exact Test.

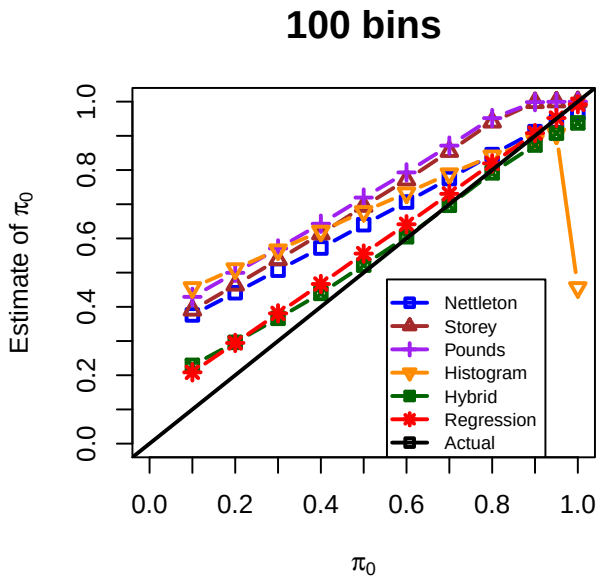


log-Normal distributions

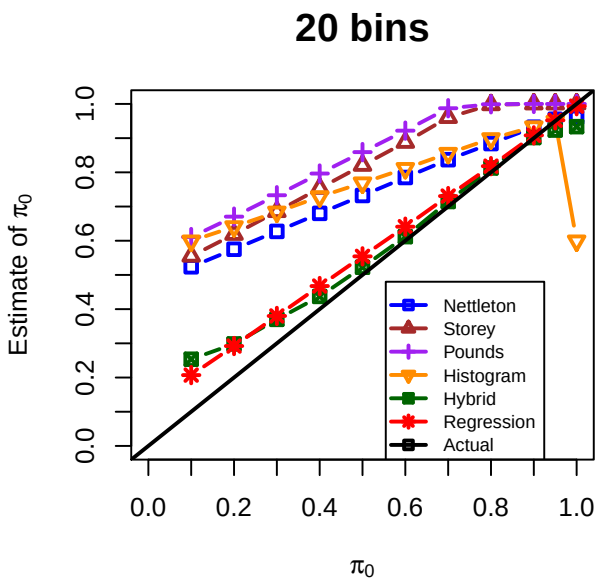
of total reads/feature

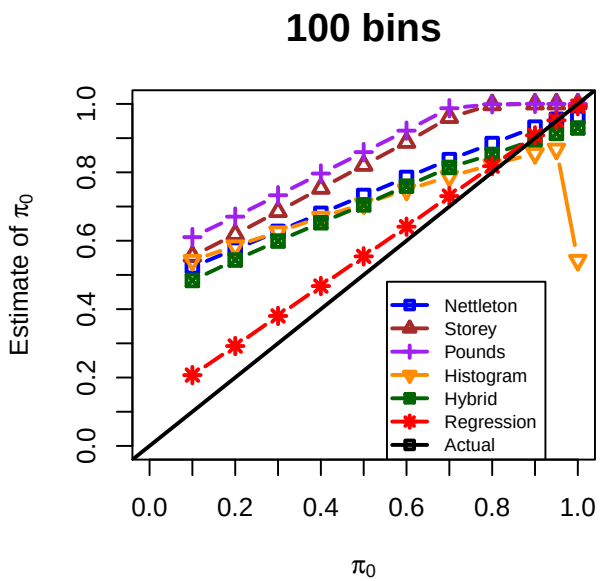
Estimated π_0 with m=10,000





logNormal (red)





logNormal (black)

Estimating FDR

- Benjamini and Hochberg (1995) suggest an algorithm for controlling FDR at level q:
  - Find the maximal i such that $p_{(i)} \leq \frac{i \times q}{m}$.
  - It has been shown when the test statistics are continuous and independent, then this algorithm controls the FDR at level $\pi_0 q$

PENNSTATE

Estimating FDR

- Benjamini and Hochberg (1995) suggest an algorithm for controlling FDR at level q:
  - Find the maximal i such that $p_{(i)} \leq \frac{i \times q}{m}$.
  - It has been shown when the test statistics are continuous and independent, then this algorithm controls the FDR at level $\pi_0 q$
- Adaptive methods take π_0 into account when setting the rejection region for the tests.

PENNSTATE

Estimating FDR

- Benjamini and Hochberg (1995) suggest an algorithm for controlling FDR at level q:
  - Find the maximal i such that $p_{(i)} \leq \frac{i \times q}{m}$.
  - It has been shown when the test statistics are continuous and independent, then this algorithm controls the FDR at level $\pi_0 q$
- Adaptive methods take π_0 into account when setting the rejection region for the tests.
- The BH method is known to be conservative with discrete tests.

PENNSTATE

Estimating FDR

- Gilbert (2005) filters tests which have zero power to achieve significance at level α and then applies the BH method.

Estimating FDR

- Gilbert (2005) filters tests which have zero power to achieve significance at level α and then applies the BH method.
- We suggest an adaptive Gilbert method that uses an estimate of π_0 with Gilbert's method.

PENNSTATE

Estimating FDR

- Gilbert (2005) filters tests which have zero power to achieve significance at level α and then applies the BH method.
- We suggest an adaptive Gilbert method that uses an estimate of π_0 with Gilbert's method.
- Prescreening will depend on value of the ancillary statistic, and so is independent of the outcome of the test.
- For count data, the ancillary statistic is the total count per feature.

PENNSTATE

Adaptive FDR

Adaptive methods:

Adaptive methods control at level q essentially by "targeting" $q/\hat{\pi}_0$.

- For π_0 near 1, this provides slightly more power and still usually rejects at $p(i) < q$.
  - i.e. it still behaves like an "adjusted p-value"
- For $pi_0 < 0.5$ we often find that $p(i) > q$.
- We do not suggest adaptive methods when $\hat{\pi}_0$ is small.
- Instead:
  - pick a rejection level α suitable for a single test.
  - estimate FDR for that level (using $\hat{\pi}_0$).

PENNSTATE

Many thanks

Thanks for your attention

PENNSTATE

Many thanks

Thanks for your attention

Thanks to NSF

- NSF DMS 1007801 (Altman, PI)
- NSF IOS 0820729 (Altman, subcontract from McSteen, PI)

Main Reference:

- Dialsingh, I (2011) False Discovery Rates when the Statistics are Discrete. PhD Dissertation, Dept. of Statistics, Penn State University

References

- 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 Series B*, 57, 289-300.
- Benjamini, Y., Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Behavioral Educational Statistics*, 25, 60-83.
- Benjamini, Y., Krieger, A.M., Yekutieli, D. (2006). *Adaptive Linear Step-up False Discovery Rate controlling procedures*. Biometrika, 93:491-507.
- Gilbert,P.B. (2005). A modified false discovery rate multiple comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of Applied Statistics*, 54, 143-158.
- Nettleton, D., Hwang, J.T.G., Caldo, R.A., Wise, R.P. (2006). Estimating the number of true null hypotheses from a histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 337-356.
- Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics*, 20, 1737-1745.

PENNSTATE