False Discovery Rates and Copy Number Variation

Bradley Efron and Nancy Zhang

Stanford University

Three Statistical Centuries

- **19th** (*Quetelet*) Huge data sets, simple questions
- **20th** (*Fisher, Neyman, Hotelling, . . .*) Small data sets, simple questions
- **21st** (*Scientific mass production*) Huge data sets, complicated questions

Example: Copy Number Variation

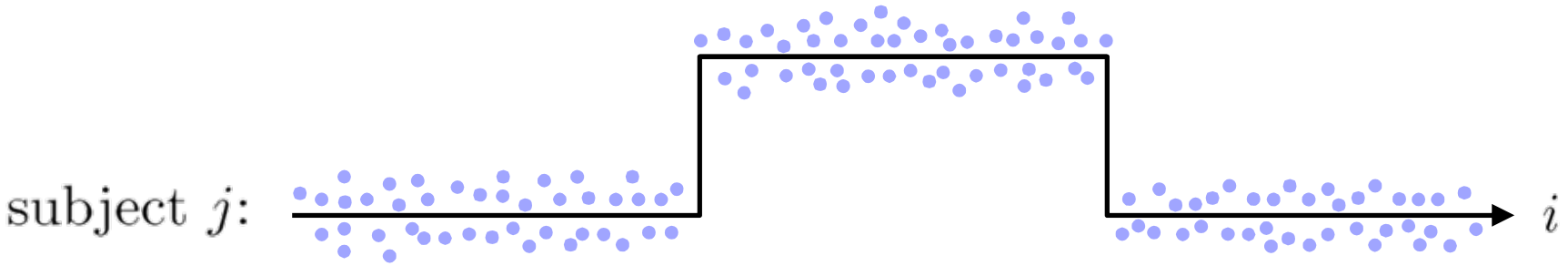
- *CNV* Gains and losses of chromosome segments (disease association)
- Instead of 2 copies, might have 0, 1, 3, 4,...
- Data

  x_{ij} = noisy msmnt of copy number for subject j at marker position i

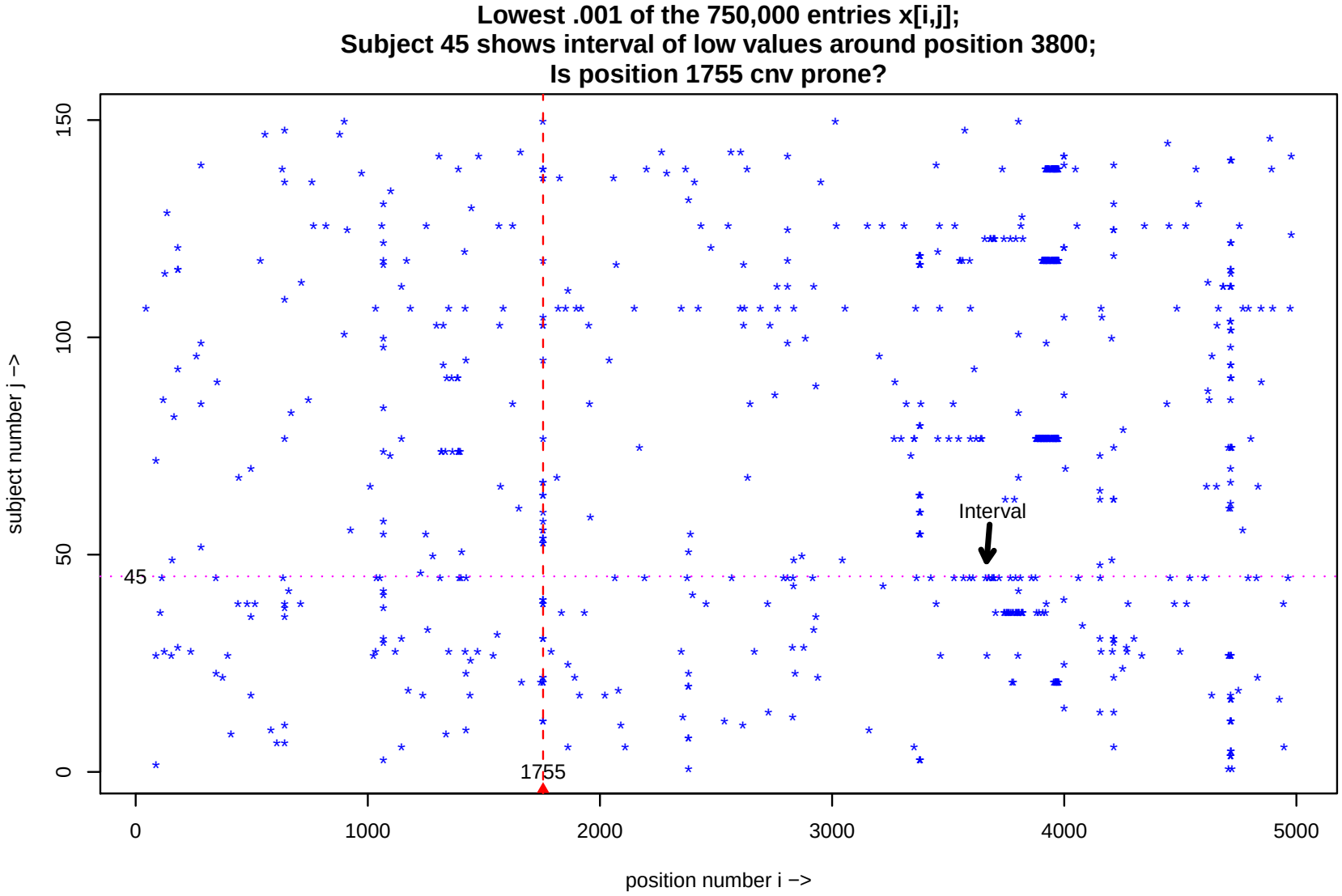
- $i = 1, 2, \ldots, N$ (5000) and $j = 1, 2, \ldots, n$ (150) ($< 1\%$ of data!)
- x_{ij} approx. normal with mean 0 if copy number = 2

What We Expect To See

- *Hot positions*: i where several of subjects j show unusually high (or low) values x_{ij}
- For some subjects j: intervals of high (or low) values x_{ij}



- *Information* on CNV locations in both directions


Lowest .001 of the 750,000 entries x[i,j];
Subject 45 shows interval of low values around position 3800;
Is position 1755 cnv prone?
subject number j –>
position number i –>
0
50
100
150
0
1000
2000
3000
4000
5000
45
1755
Interval

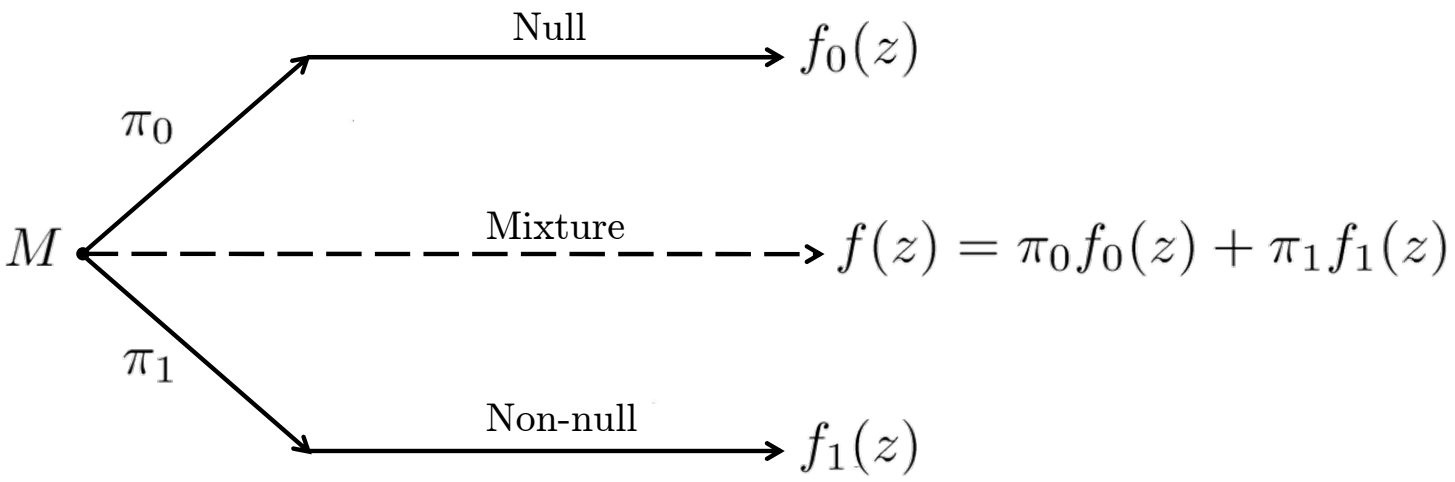
Z-Values

- $\boldsymbol{X}$: x_{ij} for $i = 1, 2, \ldots, N = 5000$ positions,
  $j = 1, 2, \ldots, n = 150$ subjects
- C_i = all subjects at ith position $\quad (n = 150)$
- Moving averages Replace x_{ij} with $\bar{x}_{ij} = \sum_{i-5}^{i+5} x_{lj}/11$
  $\Rightarrow \bar{\boldsymbol{X}}$ (j's msmnts averaged over nearby positions)
- *Standardize rows of* $\bar{\boldsymbol{X}}$
  - subtract row median
  - divide by row robust standardization
- Gives $\boldsymbol{Z}$ matrix $z_{ij} \Rightarrow$ iterative $\widehat{\text{fdr}}_i$ estimates

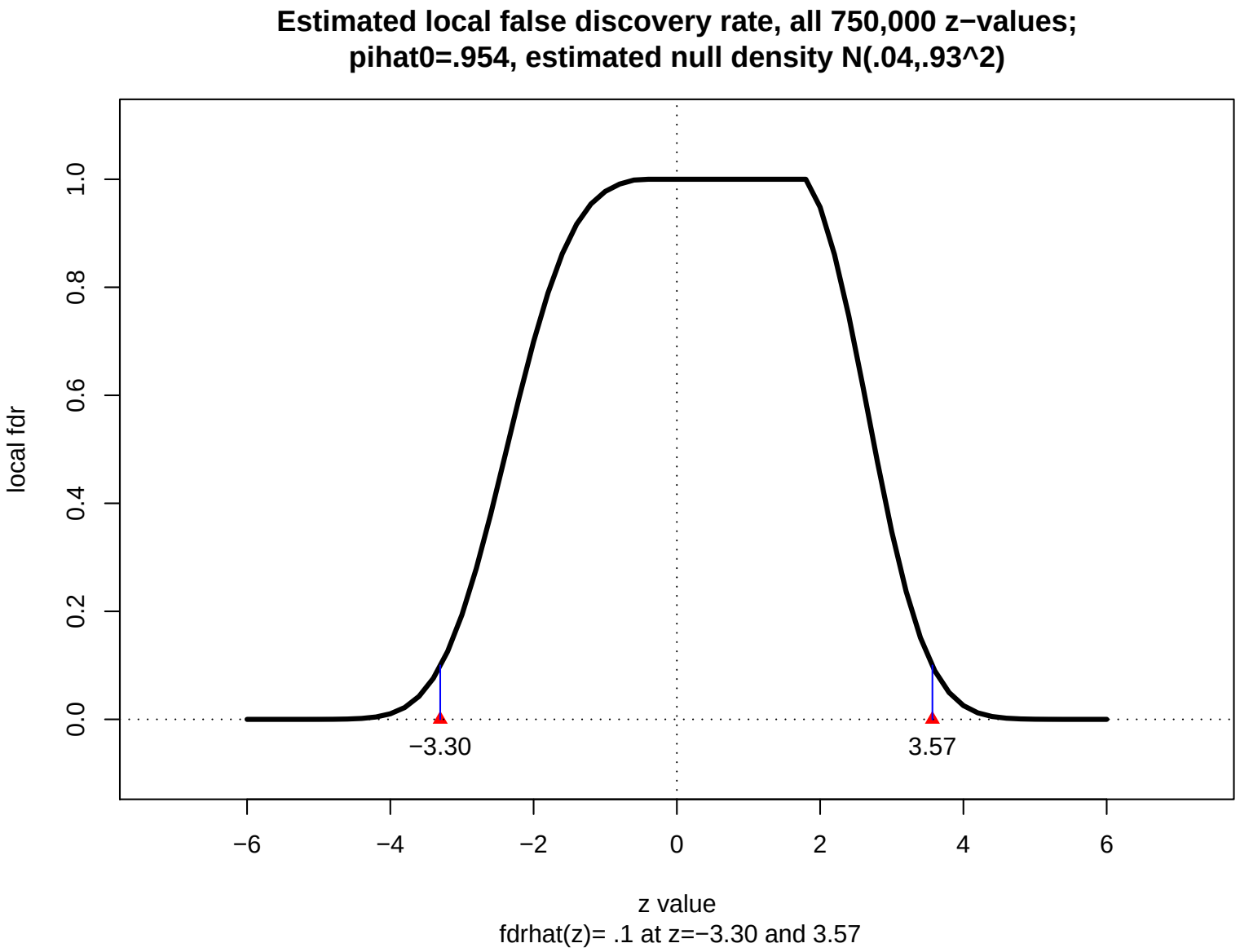
Simultaneous Hypothesis Testing

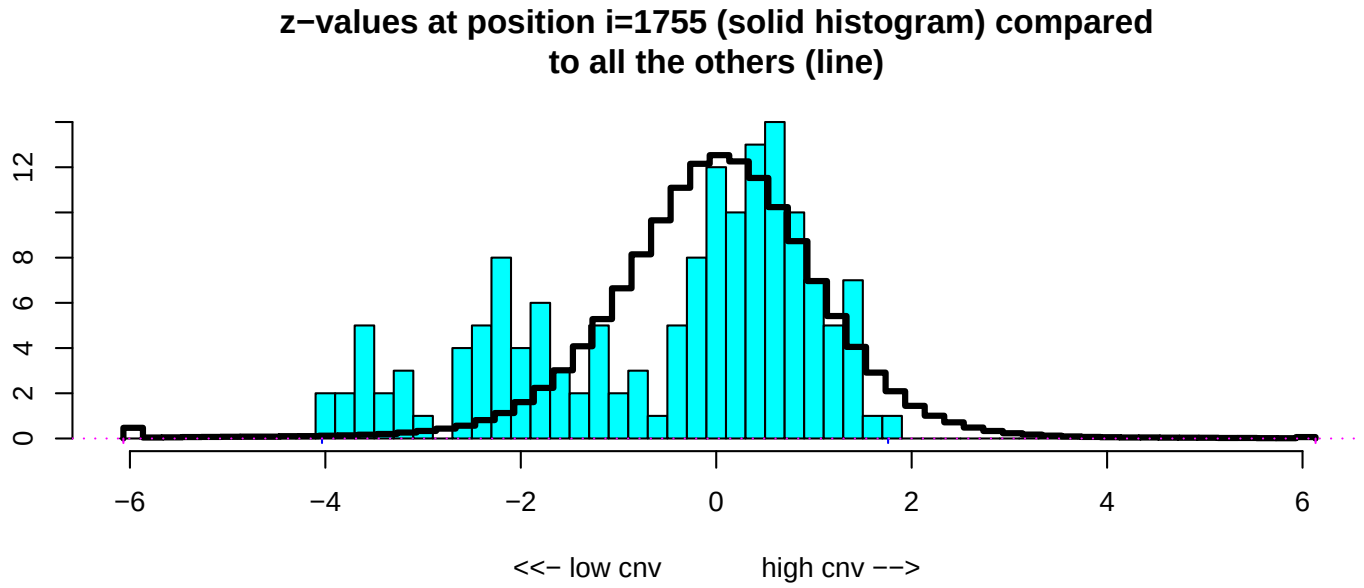
- *M null hypotheses* $H_{01}, H_{02}, \ldots, H_{0M}$
  ($M = 750,000$ for CNV)
- Case m has test statistic z_m, null density $f_0(z)$
- The problem Given $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$,
  simultaneously test all M null hypotheses — and don't make many mistakes!

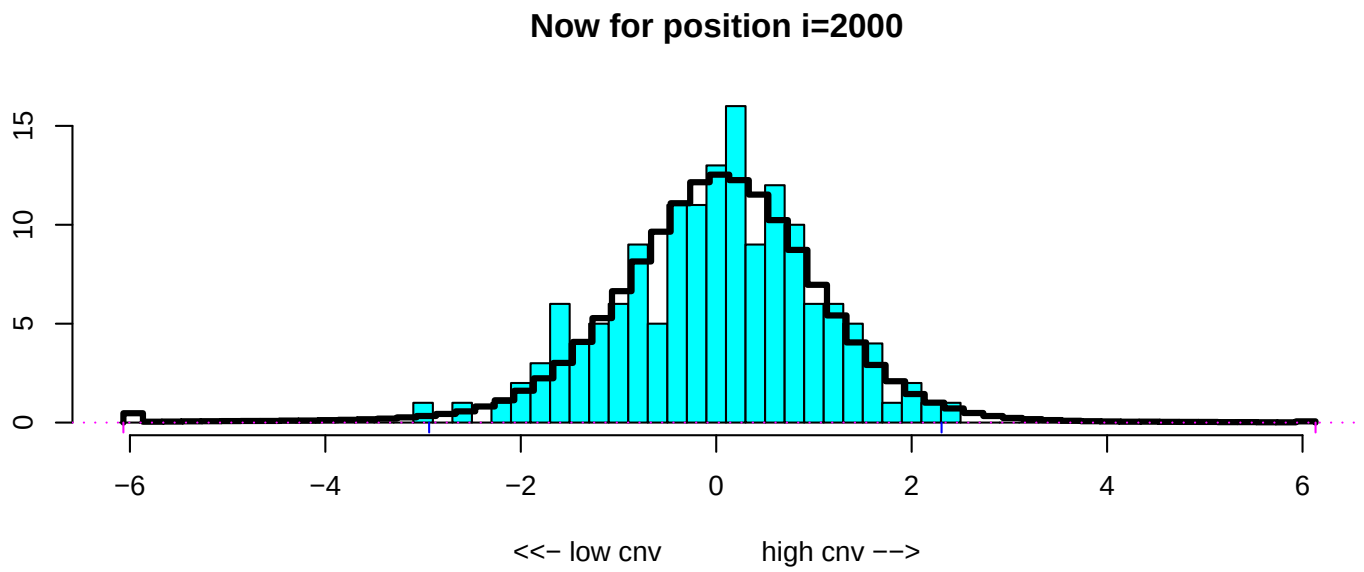
The Bayesian Two-Groups Model



- Local false discovery rate
  $\text{fdr}(z) = \Pr\{\text{null}|z\} = \pi_o f_0(z)/f(z)$
- *Empirical Bayes* $z \longrightarrow \hat{\pi}_0, \hat{f}_0, \hat{f} \longrightarrow \boxed{\widehat{\text{fdr}}(z) = \hat{\pi}_0 \hat{f}_0(z) \Big/ \hat{f}(z)}$
- Reject H_{0m} if $\widehat{\text{fdr}}(z_m)$ small (see Efron, 2008)

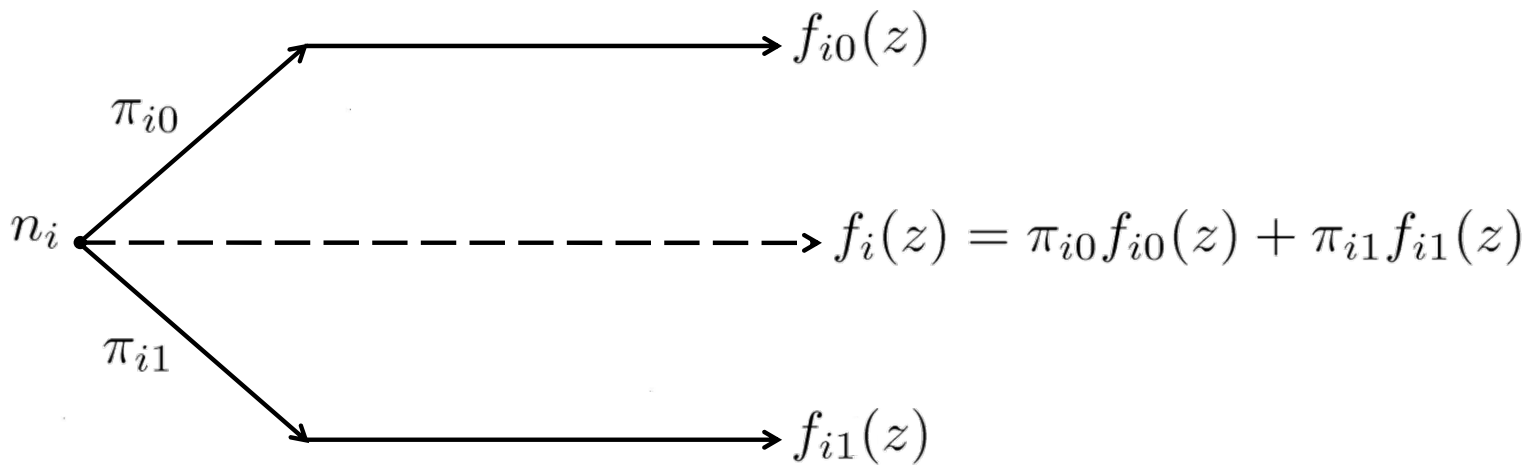

Estimated local false discovery rate, all 750,000 z−values;
pihat0=.954, estimated null density N(.04,.93^2)
local fdr
0.0
0.2
0.4
0.6
0.8
1.0
−3.30
3.57
−6
−4
−2
0
2
4
6
z value
fdrhat(z)= .1 at z=−3.30 and 3.57


z−values at position i=1755 (solid histogram) compared
to all the others (line)
0
2
4
6
8
12
−6
−4
−2
0
2
4
6
<<− low cnv
high cnv −−>


Now for position i=2000
0
5
10
15
−6
−4
−2
0
2
4
6
<<− low cnv
high cnv −−>

A More General Model

- *Classes*: $C_1, C_2, \ldots, C_i, \ldots, C_N$ with n_i cases in C_i
- CNV: $C_i = i$th column, $n_i = n = 150$
  (the $n = 150$ subjects measured at position i)



- $\mathrm{fdr}_i(z) = \pi_{i0} f_{i0}(z) / f_i(z) = \Pr\{\mathrm{null} | z, C_i\}$

Combined and Separate Fdr's

- *Strategy*: Estimate $\text{fdr}(z) = \pi_0 f_0(z)/f(z)$ from combined data and then modify for C_i
- Assume $f_{i0}(z)$ and $f_{i1}(z)$ do not depend on i, only $\pi_{i1} = \Pr\{\text{non-null}|C_i\}$ varying across classes:

$$\boxed{\text{fdr}_i(z) = \text{fdr}(z) \Big/ [1 + \text{tdr}(z)S_i]}$$

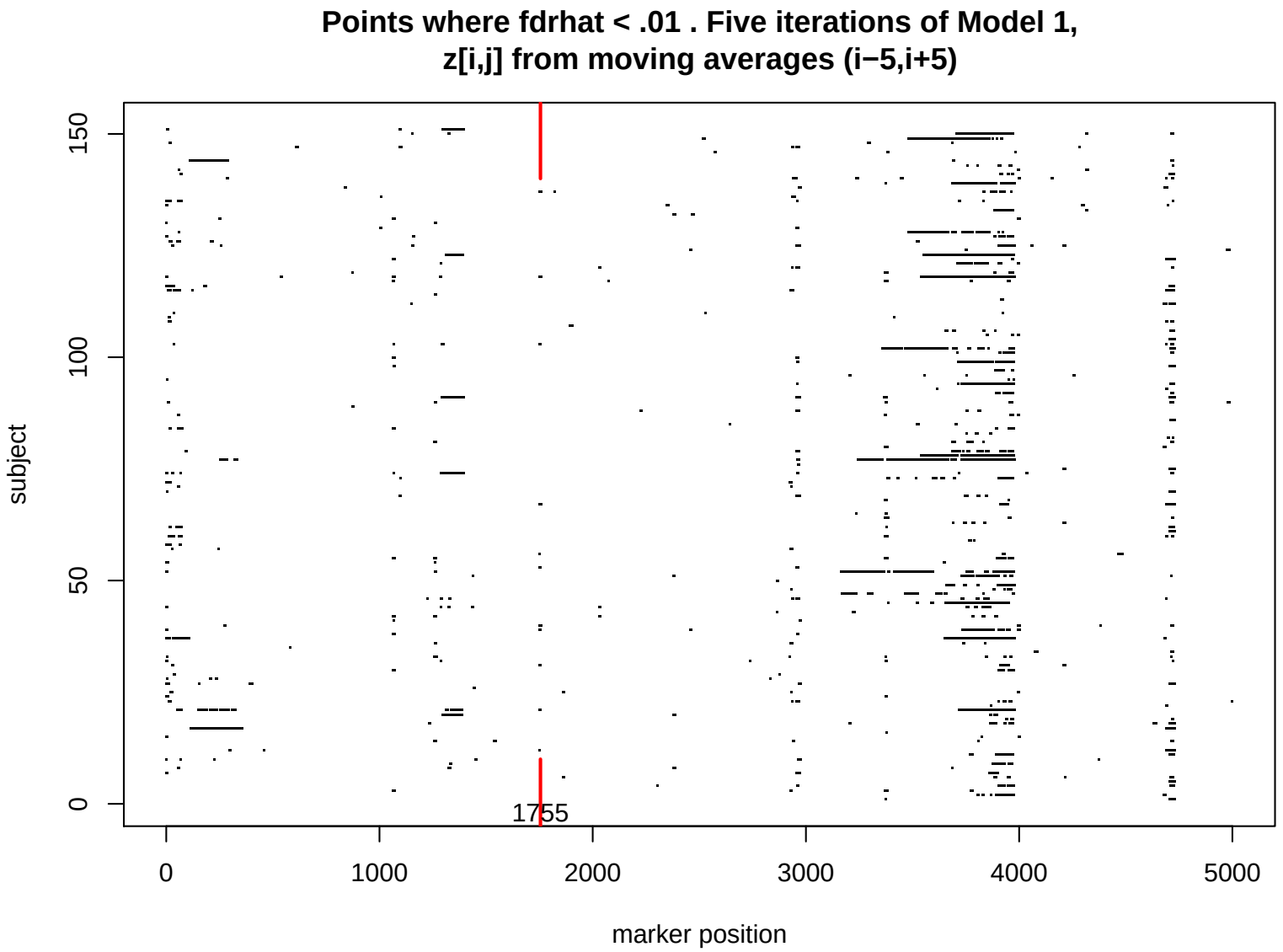
- $\text{tdr}(z) \equiv 1 - \text{fdr}(z) =$ "true discovery rate" and $S_i = \dfrac{\pi_{i1}}{\pi_1} \Big/ \dfrac{\pi_{i0}}{\pi_0} - 1$

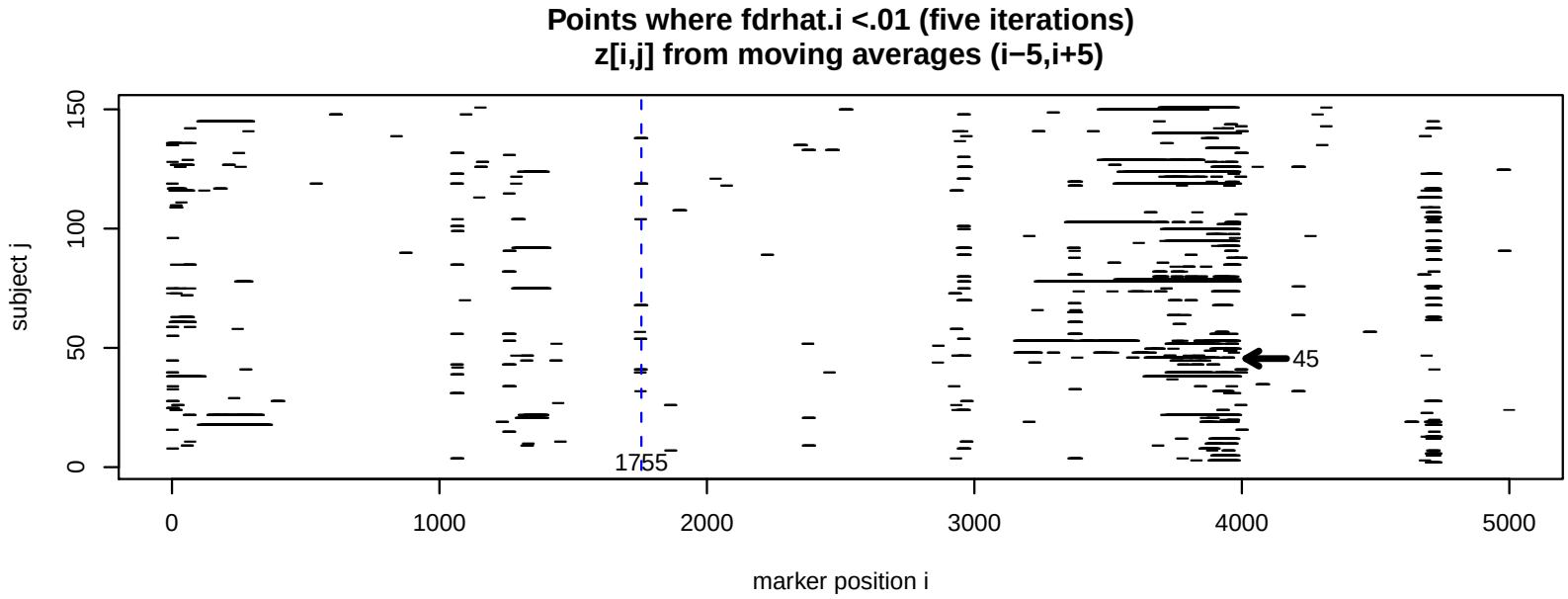
Iterative Estimation of $\mathrm{fdr}_i(z)$ (Model 1)

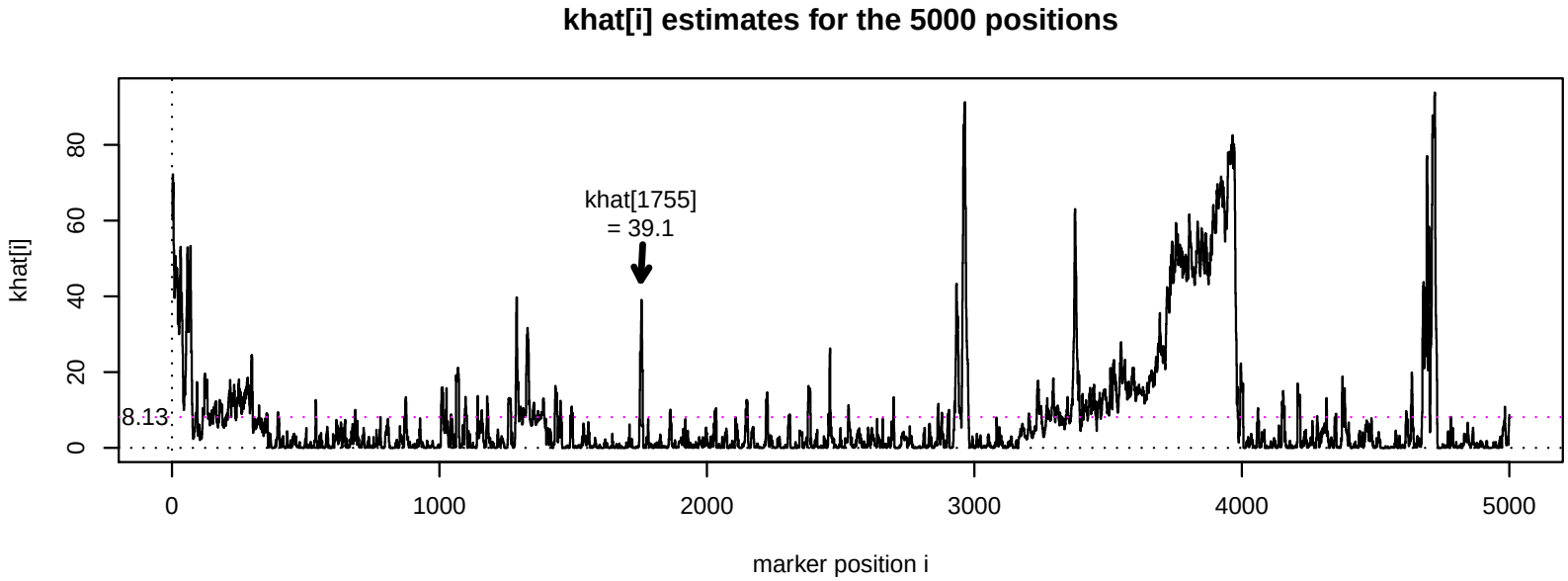
- First Estimate $\widehat{\mathrm{fdr}}(z) = \hat{\pi}_0 \hat{f}_0(z) \Big/ \hat{f}(z)$ from combined data $(z_1, z_2, \ldots, z_M)$
- If k_i non-nulls in C_i: $\hat{\pi}_{i1} = k_i/n$ gives $\hat{S}_i$ and

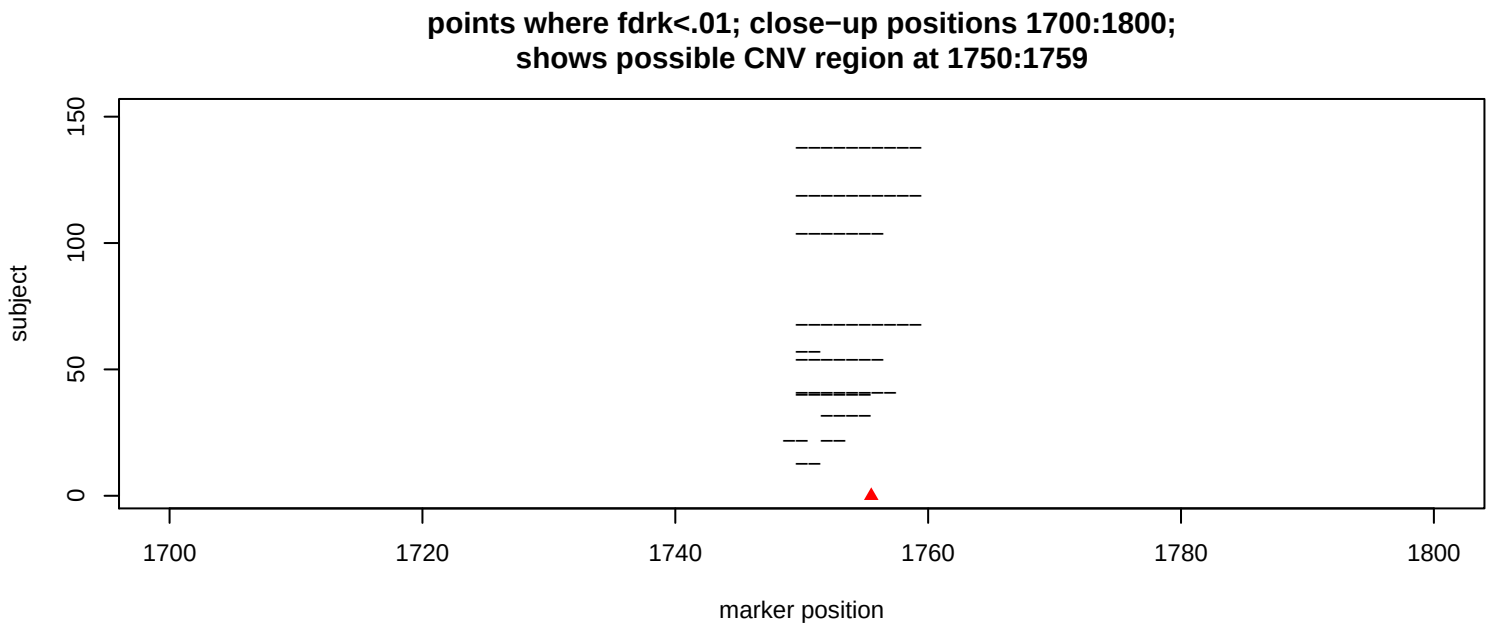
$$\widehat{\mathrm{fdr}}_i(z) = \frac{\widehat{\mathrm{fdr}}(z)}{1 + \widehat{\mathrm{tdr}}(z)\hat{S}_i} \qquad \left(\widehat{\mathrm{tdr}}_i = 1 - \widehat{\mathrm{fdr}}_i\right)$$

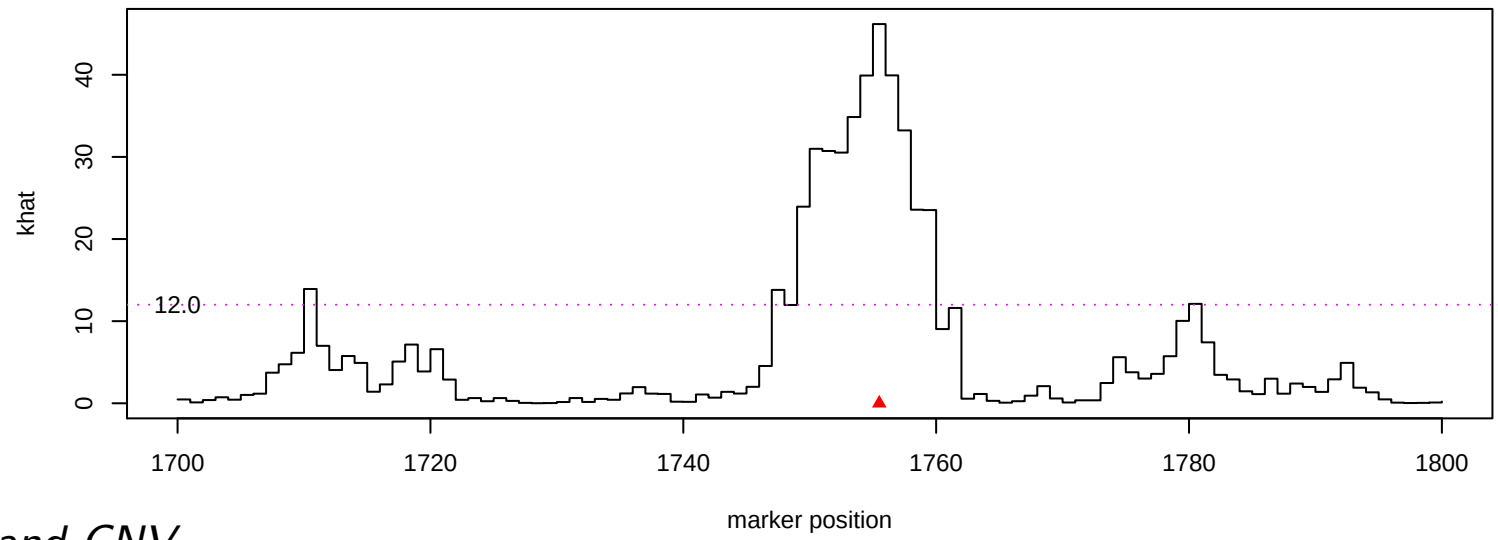
- But $\boxed{\hat{k}_i = \sum_{C_i} \widehat{\mathrm{tdr}}_i(z_m)}$ estimates k_i
- Iterate! (5 cycles plenty in what follows)


Points where fdrhat < .01 . Five iterations of Model 1,
z[i,j] from moving averages (i−5,i+5)
subject
0
50
100
150
1755
0
1000
2000
3000
4000
5000
marker position


Points where fdrhat.i <.01 (five iterations)
z[i,j] from moving averages (i−5,i+5)
subject j
0
50
100
150
1755
45
0
1000
2000
3000
4000
5000
marker position i


khat[i] estimates for the 5000 positions
khat[i]
0
20
40
60
80
khat[1755]
= 39.1
8.13
0
1000
2000
3000
4000
5000
marker position i


points where fdrk<.01; close−up positions 1700:1800;
shows possible CNV region at 1750:1759
subject
0
50
100
150
1700
1720
1740
1760
1780
1800
marker position

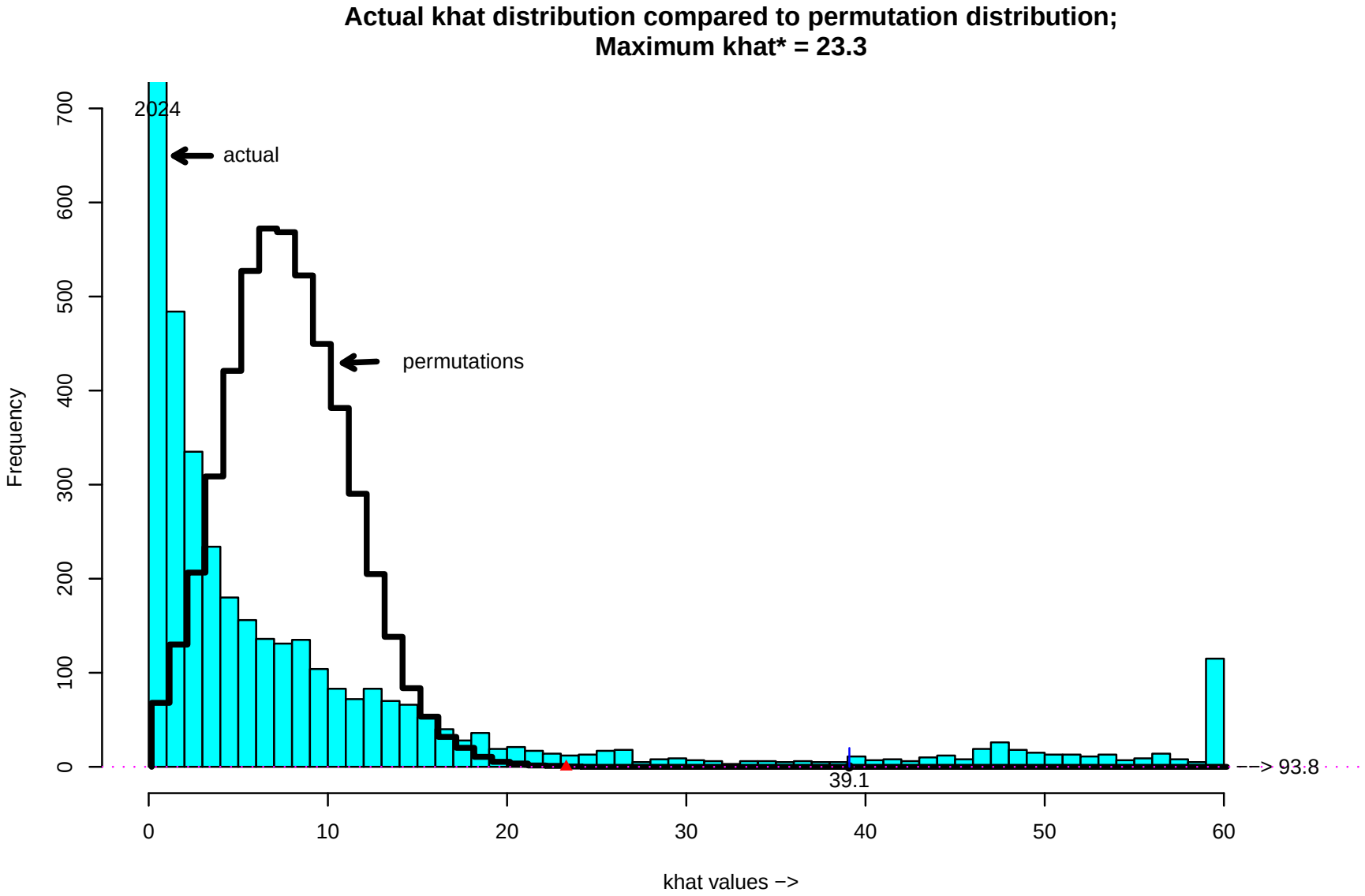

khat
0
10
20
30
40
12.0
1700
1720
1740
1760
1780
1800
marker position

Is Position 1755 "Significant"?

- $\hat{k}_{1755} = 39.1$ • Believe CNV action at 1755? $[\bar{k} = 8.13]$
- Permutation test Randomly shift row j of $\boldsymbol{X}$ by s_j units left (with wraparound):

$$\boldsymbol{x}_j^* = (x_{s+1,j}, x_{s+2,j}, \ldots, x_{5000,j}, x_{1j}, x_{2j}, \ldots, x_{sj})$$

- Do this for all 150 rows
- *Recalculate* $\hat{k}_i^*$ values
- Compare $\hat{k}_{1755} = 39.1$ with $\left\{\hat{k}_i^*,\ i = 1, 2, \ldots, 5000\right\}$


Actual khat distribution compared to permutation distribution;
Maximum khat* = 23.3
2024
actual
permutations
Frequency
0
100
200
300
400
500
600
700
39.1
--> 93.8
0
10
20
30
40
50
60
khat values ->

Locally Most Powerful Tests

- Let $r_i = \pi_{i1}/\pi_1 = \Pr\{\text{non-null}|C_i\} \Big/ \Pr\{\text{non-null}\}$.
- $l_i = \sum_1^n \left\{1 + (r_i - 1)T(z_{ij})\right\}$ where $T(z) = \dfrac{\text{tdr}(z) - \pi_1}{\pi_0}$
- $\hat{k}_i$ nearly MLE in this model
- Test

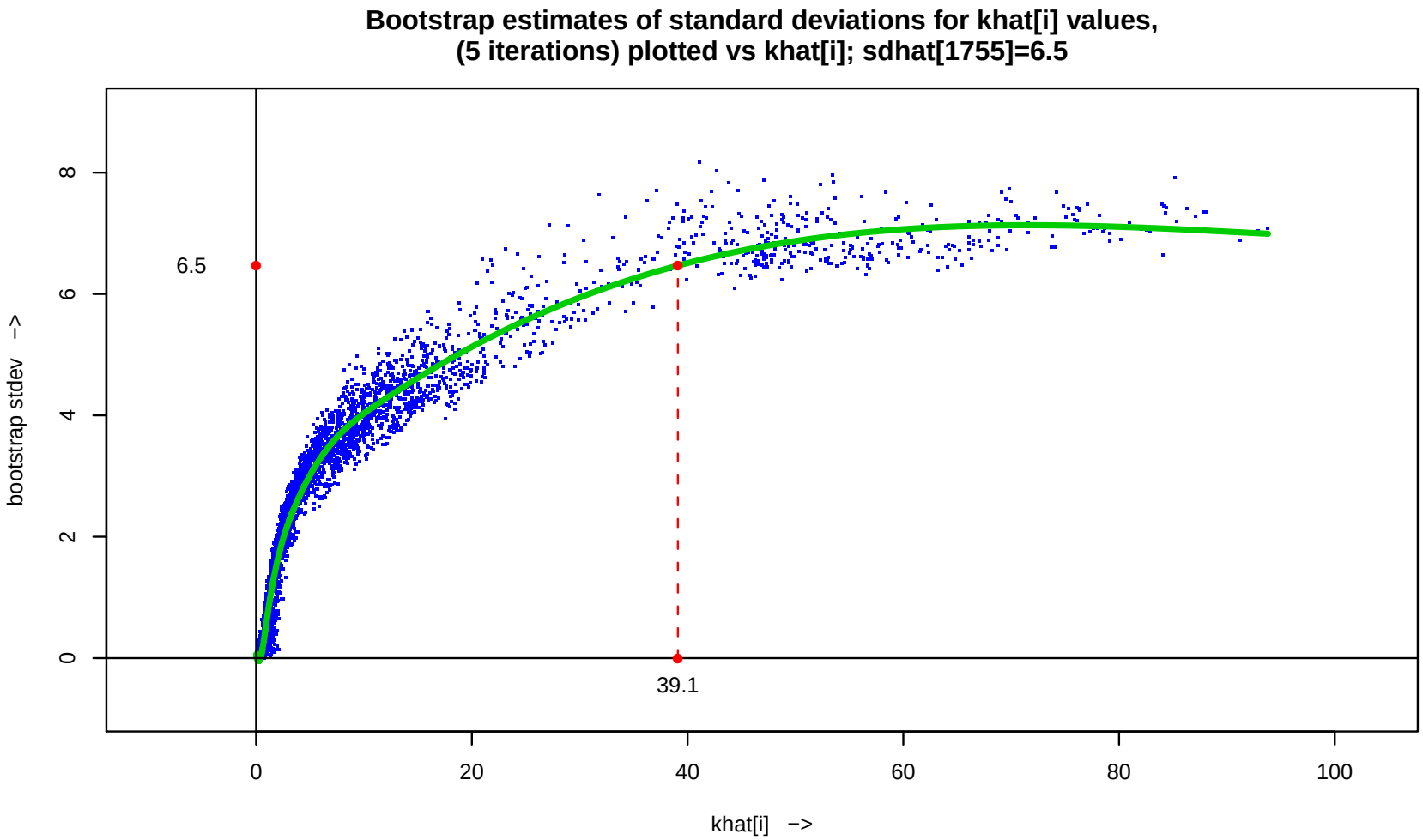
$$H_{0i} : r_i = 1 \quad \text{vs} \quad r_i > 1.$$

- Locally most powerful test rejects for large values of $\hat{k}_i^{(1)}$.
- Use permutations to get p-values.

Bootstrapping $\hat{k}_i$ Estimates

- Resample rows (i.e., subjects)
- Recompute iterative estimate $\hat{k}_i^*$ (5 iterations model)
- $\widehat{\text{sd}}_i$ = boot stdev of $\hat{k}_i^*$, $B = 100$ resamples

  (did not recompute original $\widehat{\text{fdr}}$ curve each time)
- $\hat{k}_i^* \dot{\sim} \mathcal{N}\left(\hat{k}_i, \widehat{\text{sd}}_i^2\right)$ $\quad \left[6 \le \widehat{\text{sd}}_i \le 7 \text{ for } \hat{k}_i > 20\right]$


Bootstrap estimates of standard deviations for khat[i] values,
(5 iterations) plotted vs khat[i]; sdhat[1755]=6.5
bootstrap stdev –>
khat[i] –>
6.5
39.1
0
2
4
6
8
0
20
40
60
80
100

Brown–Stein–Robbins Estimation

- Suppose $\mu \sim g(\cdot)$ and $x|\mu \sim \mathcal{N}(\mu, \sigma^2)$
- $l(x)$ log marginal density of x

$$\boxed{\mu|x \sim (x + \sigma^2 l'(x)),\ \sigma^2 (1 + \sigma^2 l''(x))}$$

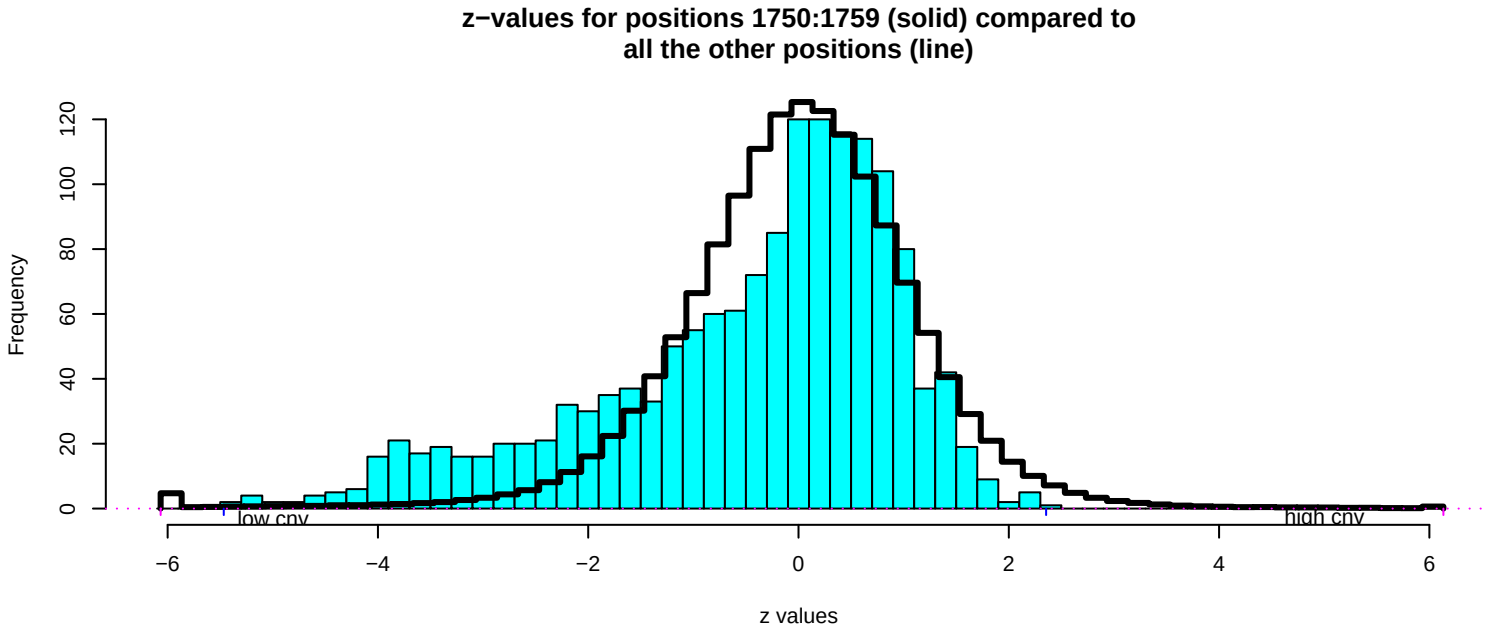
- Apply with $\mu = k_i,\ x = \hat{k}_i$,
  $\hat{l}(x) =$ log smoothed density $\left\{\hat{k}_i\right\}$
- For $\hat{k}_i = 39.1,\ \hat{\sigma} = 6.5$, gave $k_{1755} \dot{\sim} (41.3, 7.4^2)$
- Conclusion Even taking account of selection effects, k_{1755} is probably much larger than $\bar{k} = 8.13$.

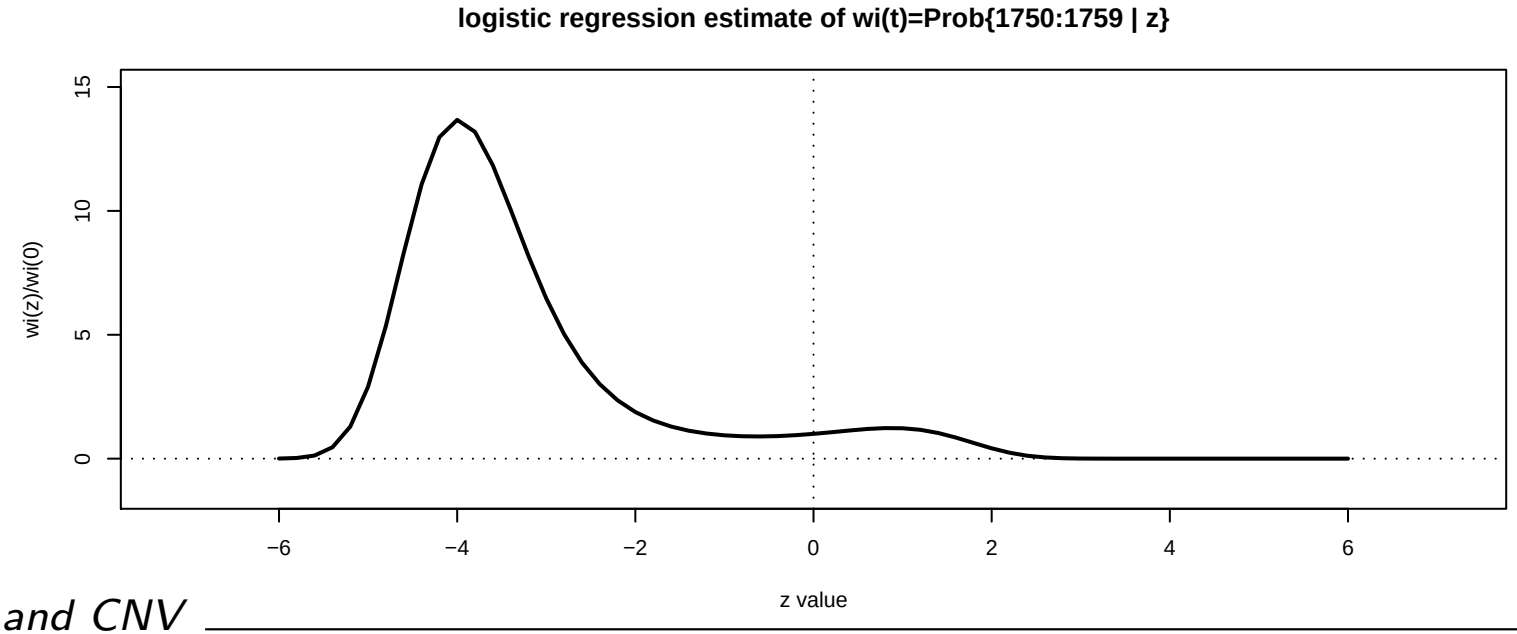
More General Model for $\text{fdr}_i(z)$

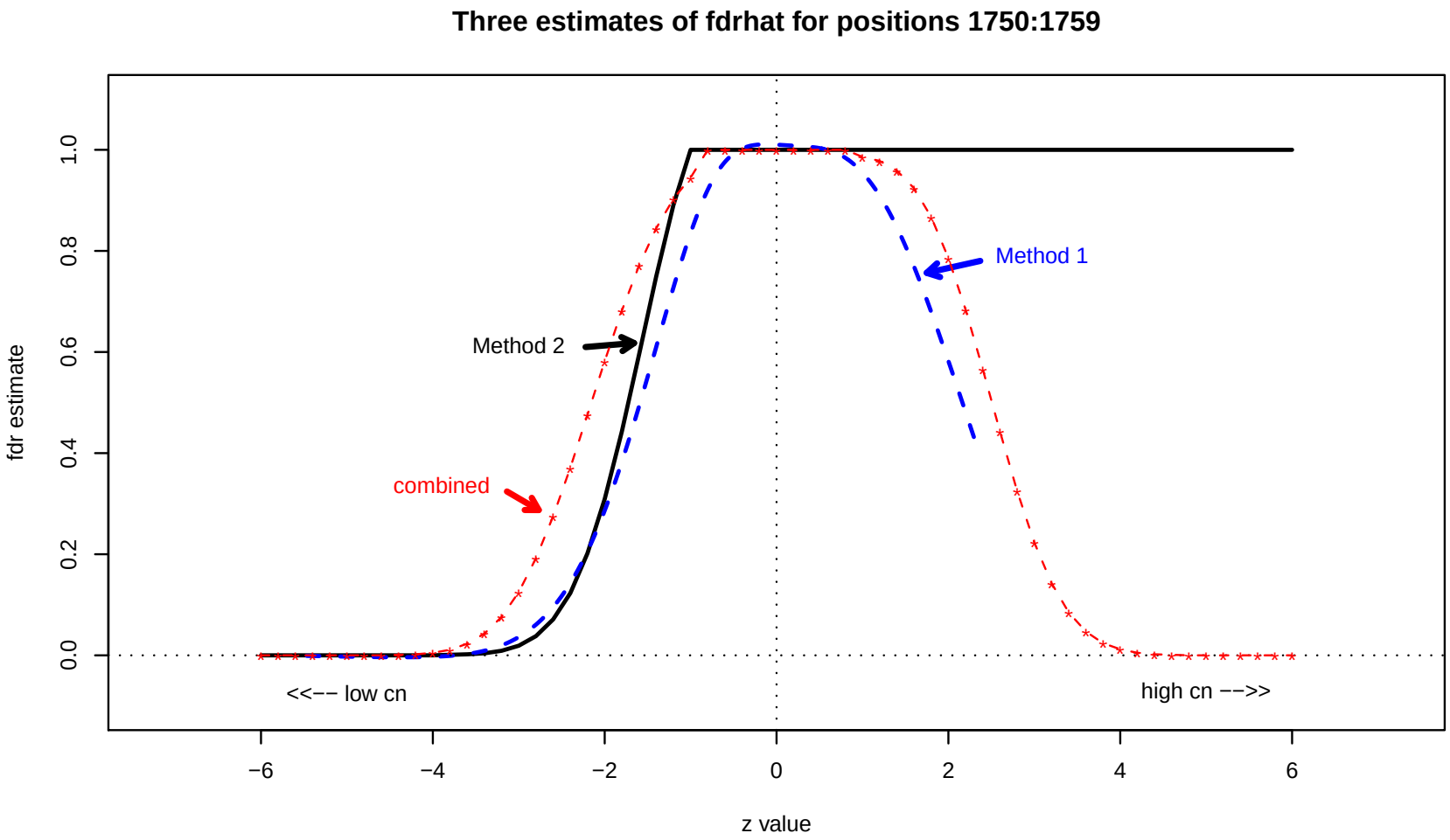
- "Method 2": Multiclass Bayes model π_{i0}, $f_{i0}(z)$, $f_{i1}(z)$ with all $f_{i0} = f_0$, but *drop assumption* that non-null distributions $f_{i1}(z)$ the same.
- *Define*: $w_i(z) = \Pr\{C_i|z\}$

$$\text{fdr}_i(z) \doteq \text{fdr}(z) \cdot \frac{w_i(0)}{w_i(z)}$$

- Empirical Bayes Estimate $w_i(z)$ by logistic regression of C_i indicator on z_m.


z−values for positions 1750:1759 (solid) compared to
all the other positions (line)
Frequency
0
20
40
60
80
100
120
low cnv
high cnv
−6
−4
−2
0
2
4
6
z values


logistic regression estimate of wi(t)=Prob{1750:1759 | z}
wi(z)/wi(0)
0
5
10
15
−6
−4
−2
0
2
4
6
z value


Three estimates of fdrhat for positions 1750:1759
fdr estimate
0.0
0.2
0.4
0.6
0.8
1.0
Method 2
Method 1
combined
<<−− low cn
high cn −−>>
−6
−4
−2
0
2
4
6
z value

References

Efron, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Statist.*, **2** 197–223.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, **9** 18–29.

Walther, G. (2009). Optimal and fast detection of spatial clusters with scan statistics. Online, URL `http://stat.stanford.edu/~gwalther/`.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B.

and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics*, **6** 45–58.

Zhang, N., Siegmund, D., Ji, H. and Li, J. (2009). Detecting simultaneous change-points in multiple sequences. *Biometrika*. Accepted for publication, URL `http://stat.stanford.edu/~nzhang/`.