

SciLifeLab

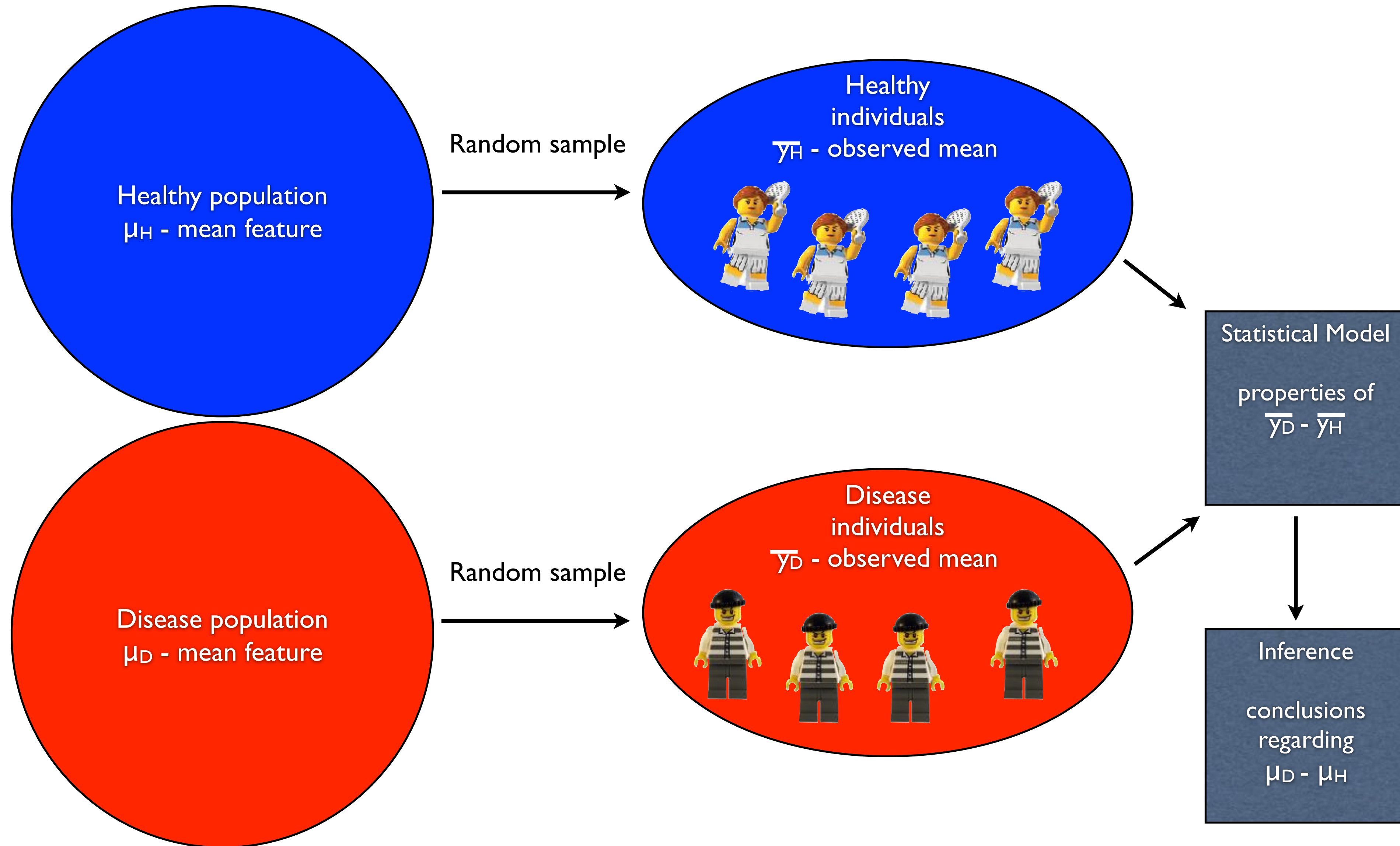
Multiple Hypothesis Testing

CB2030

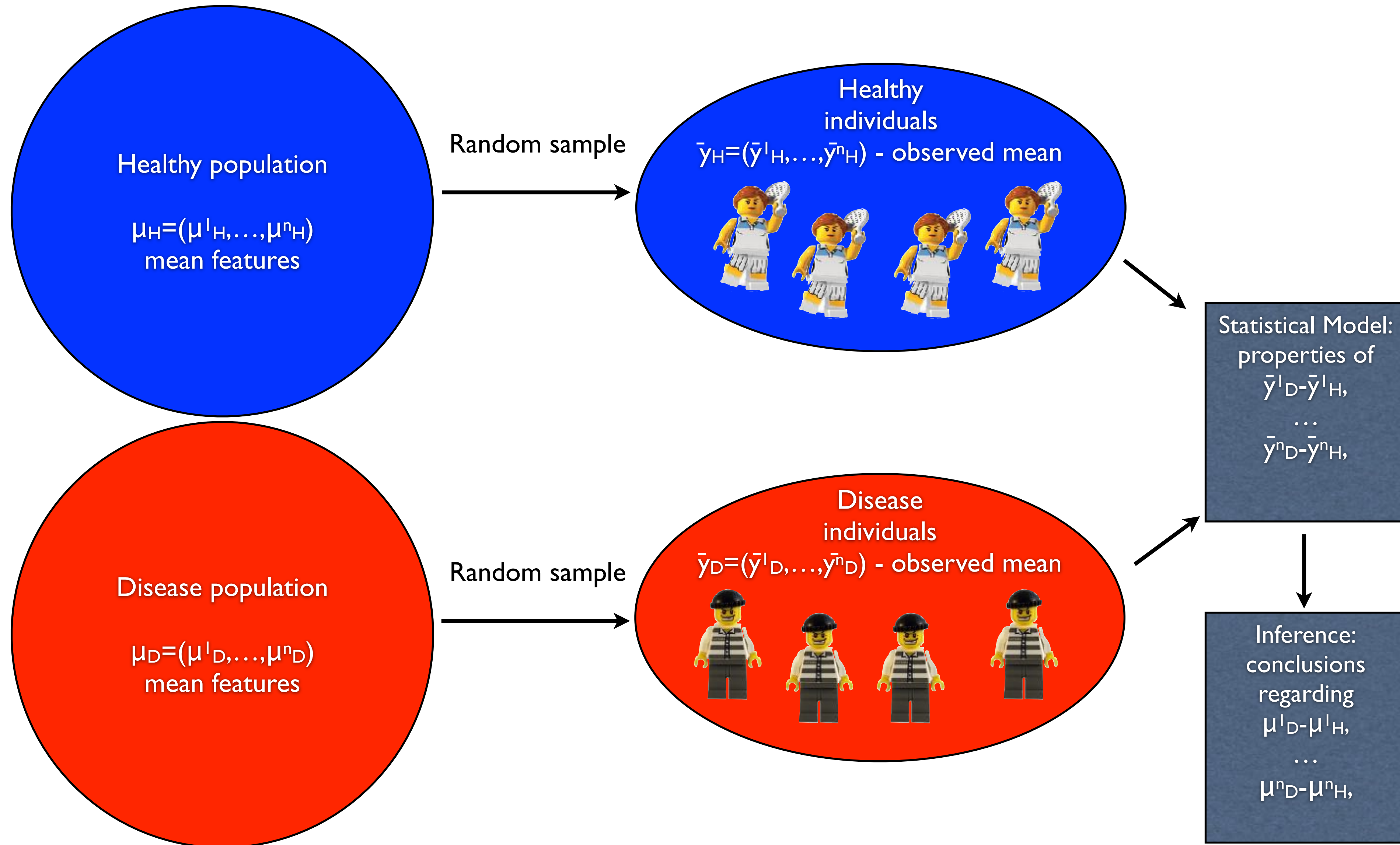
Lukas Käll, KTH



Statistical inference procedure



Multiple measurements per sampled individual



*if you think you're
one in a million,
there are six
thousand other
people exactly like
you.*

Motivating Example: micro Array study (published in Nature)

cision. Gene expression levels were compared using one-way ANOVA. This yielded 77, 642 and 2,492 differentially expressed genes at unadjusted $P < 0.001$, $P < 0.01$ and $P < 0.05$ levels, respectively. Differentially expressed genes

How many of 50 000 probes would we expect to be significant under the null hypothesis?

cision. Gene expression levels were compared using one-way ANOVA. This yielded 77, 642 and 2,492 differentially expressed genes at unadjusted $P < 0.001$, $P < 0.01$ and $P < 0.05$ levels, respectively. Differentially expressed genes

How many of 50 000 probes would we expect to be significant under the null hypothesis?

$$\text{with } P < 0.001: 50000 * 0.001 = 50$$

cision. Gene expression levels were compared using one-way ANOVA. This yielded 77, 642 and 2,492 differentially expressed genes at unadjusted $P < 0.001$, $P < 0.01$ and $P < 0.05$ levels, respectively. Differentially expressed genes

How many of 50 000 probes would we expect to be significant under the null hypothesis?

with $P < 0.001$: $50000 * 0.001 = 50$

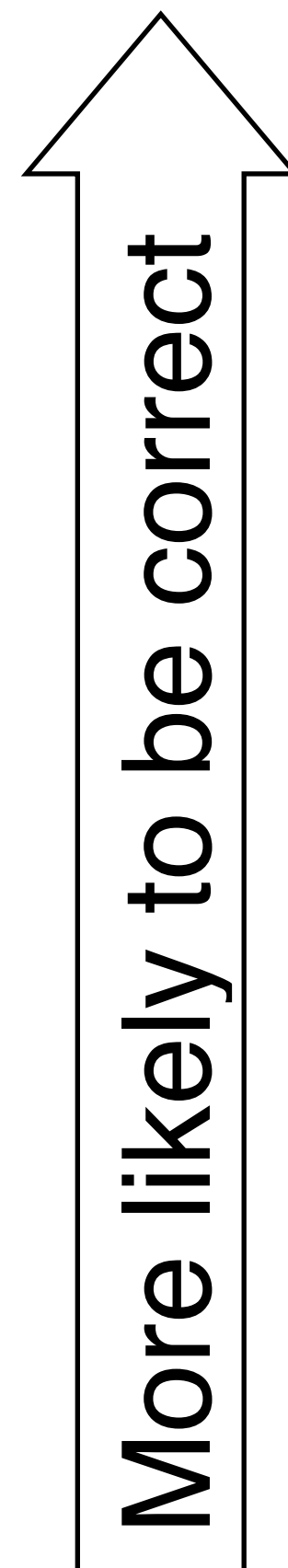
with $P < 0.01$: $50000 * 0.01 = 500$

with $P < 0.05$: $50000 * 0.05 = 2500$

cision. Gene expression levels were compared using one-way ANOVA. This yielded 77, 642 and 2,492 differentially expressed genes at unadjusted $P < 0.001$, $P < 0.01$ and $P < 0.05$ levels, respectively. Differentially expressed genes

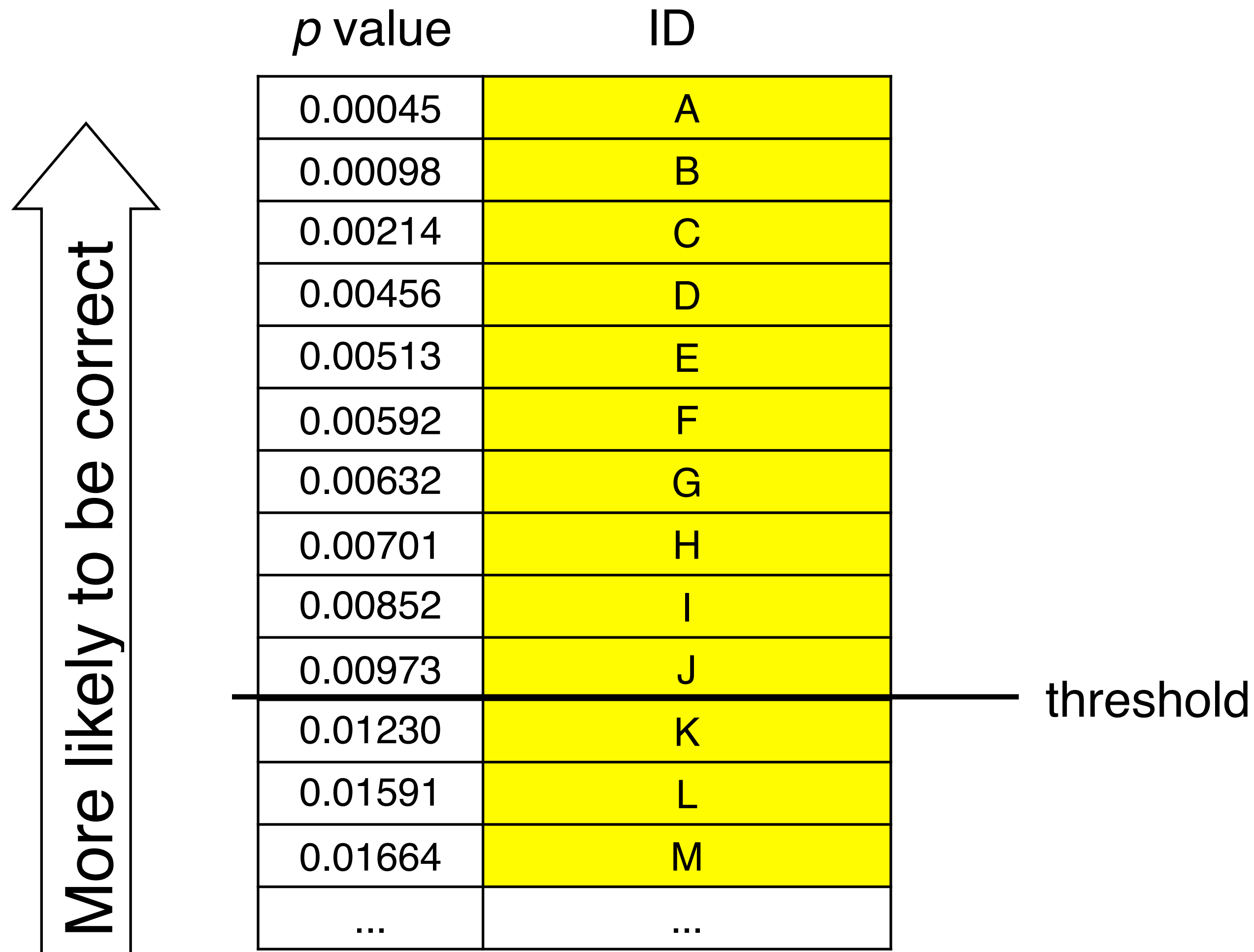
Thresholds defines which findings we report

More likely to be correct



<i>p</i> value	ID
0.00045	A
0.00098	B
0.00214	C
0.00456	D
0.00513	E
0.00592	F
0.00632	G
0.00701	H
0.00852	I
0.00973	J
0.01230	K
0.01591	L
0.01664	M
...	...

Thresholds defines which findings we report



<i>p</i> value	ID
0.00045	A
0.00098	B
0.00214	C
0.00456	D
0.00513	E
0.00592	F
0.00632	G
0.00701	H
0.00852	I
0.00973	J
0.01230	K
0.01591	L
0.01664	M
...	...

More likely to be correct

threshold

Thresholds defines which findings we report

<i>p</i> value	ID
0.00045	A
0.00098	B
0.00214	C
0.00456	D
0.00513	E
0.00592	F
0.00632	G
0.00701	H
0.00852	I
0.00973	J
0.01230	K
0.01591	L
0.01664	M
...	...

More likely to be correct

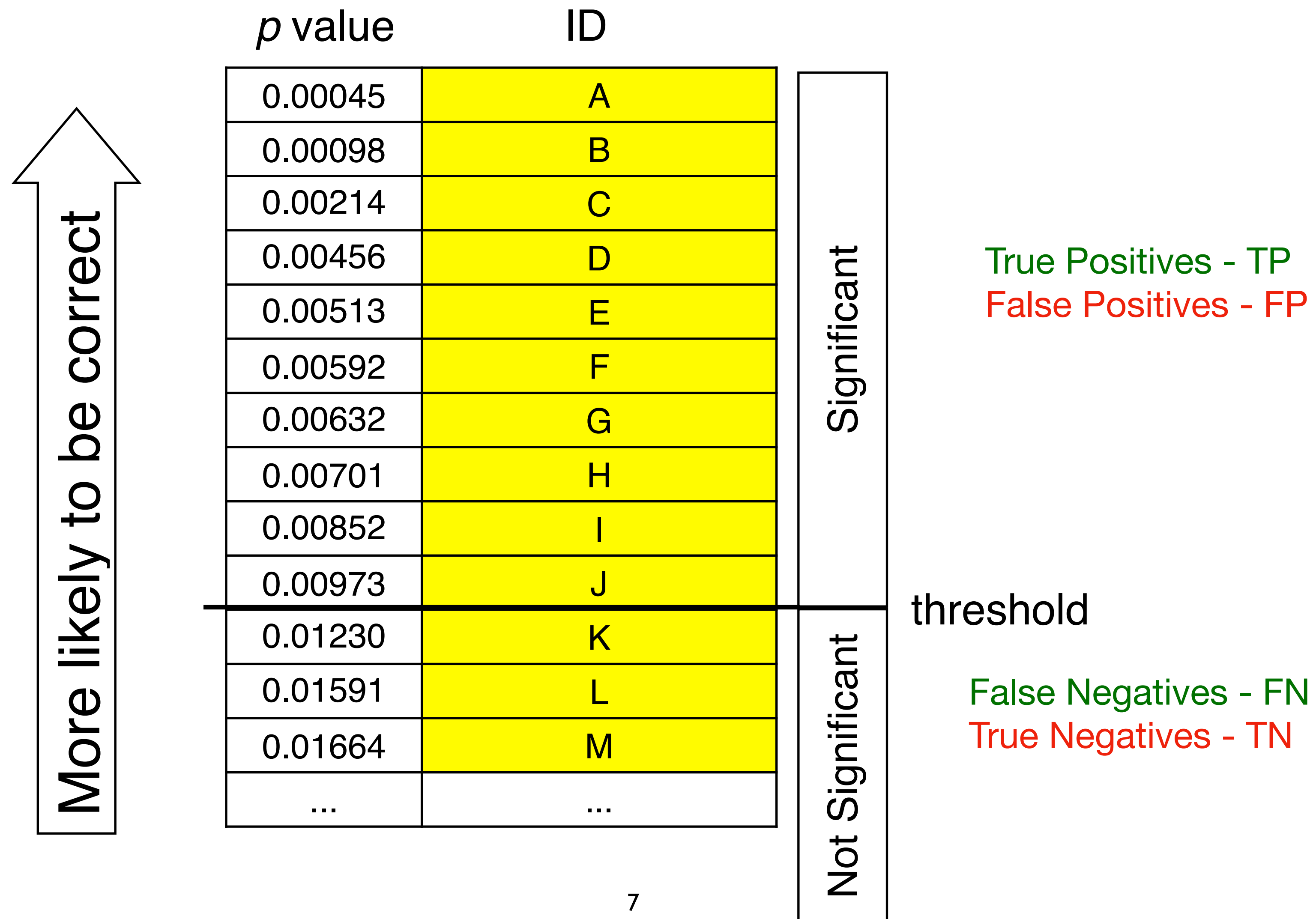
Significant

Not Significant

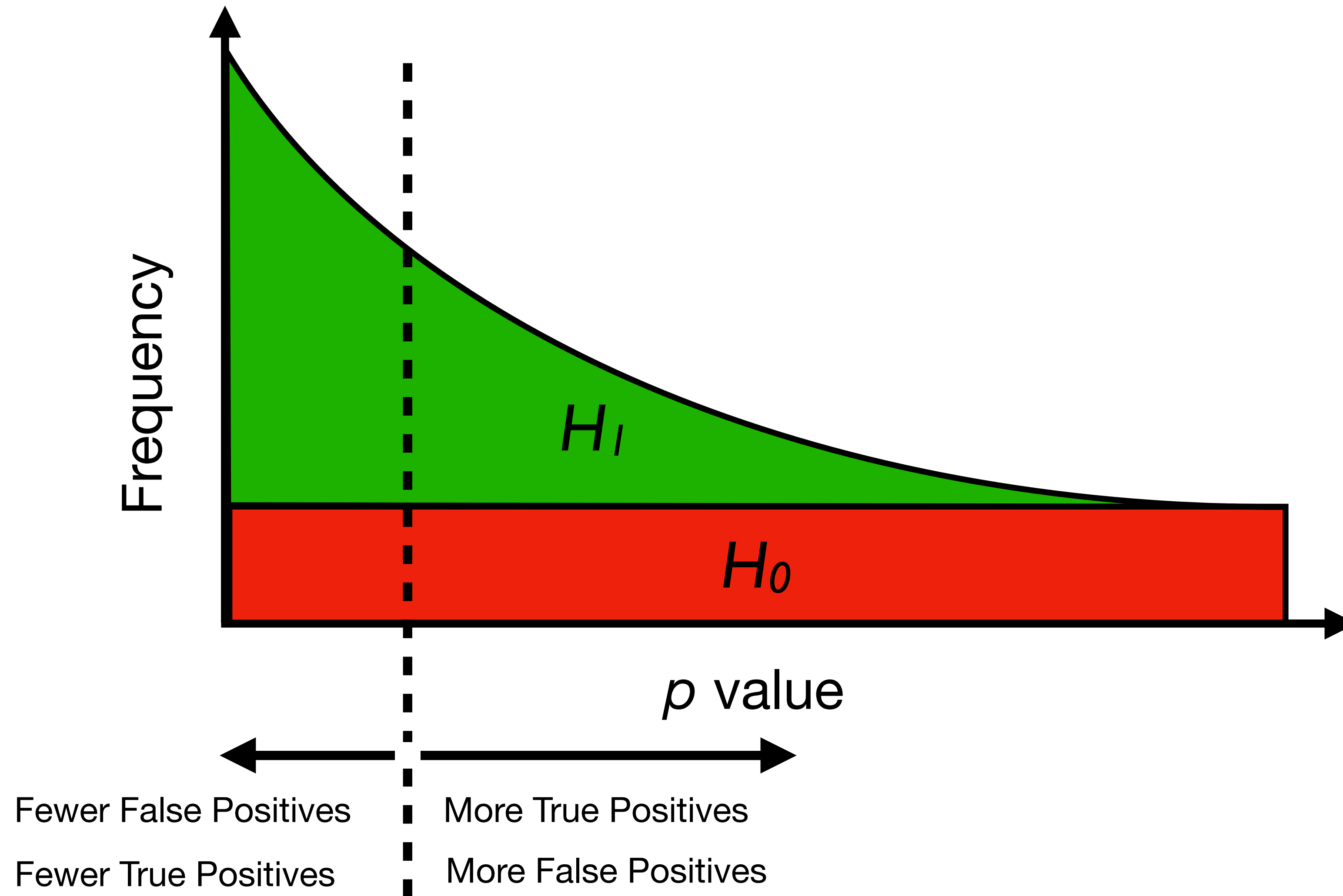
threshold

7

Thresholds defines which findings we report



How should we select threshold?



False Discovery Rate

score	type
0,0001	alternative (H_1)
0,00015	alternative (H_1)
0,00017	alternative (H_1)
0,0002	alternative (H_1)
0,00022	null (H_0)
0,00023	alternative (H_1)
0,00034	alternative (H_1)
0,00042	alternative (H_1)
0,00046	null (H_0)
0,00055	alternative (H_1)
0,00065	null (H_0)
0,00073	alternative (H_1)
0,00084	null (H_0)
...	...

threshold

False Discovery Rate

score	type
0,0001	alternative (H ₁)
0,00015	alternative (H ₁)
0,00017	alternative (H ₁)
0,0002	alternative (H ₁)
0,00022	null (H ₀)
0,00023	alternative (H ₁)
0,00034	alternative (H ₁)
0,00042	alternative (H ₁)
0,00046	null (H ₀)
0,00055	alternative (H ₁)
0,00065	null (H ₀)
0,00073	alternative (H ₁)
0,00084	null (H ₀)
...	...

$\frac{2}{10}$

threshold

$FDR(x)$ is the expectation value of the fraction of tests below threshold x that are generated under the null hypothesis

Model of differential expression

- We are studying a number of differences in feature means, some generated under the alternative hypothesis (H_1) and some to generated under the null hypothesis (H_0).



$$\Pr(p=t) = \Pr(H=H_0)\Pr(p=t|H=H_0) + \Pr(H=H_1)\Pr(p=t|H=H_1)$$

$$f(t) = \pi_0 f_0(t) + \pi_1 f_1(t)$$

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

idée [Benjamini and Hochberg 1995] - control for:

$$\frac{\text{no. false positive features}}{\text{no. significant features}} = \frac{F}{F + T} = \frac{F}{S}$$

Statistical significance for genomewide studies

John D. Storey*[†] and Robert Tibshirani[‡]

*Department of Biostatistics, University of Washington, Seattle, WA 98195; and [‡]Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved May 30, 2003 (received for review January 28, 2003)

With the increase in genomewide experiments and the sequencing of multiple genomes, the analysis of large data sets has become commonplace in biology. It is often the case that thousands of features in

to the method in ref. 5 under certain assumptions. Also, ideas similar to FDRs have appeared in the genetics literature (1, 13).

Similarly to the p value, the q value gives each feature its own

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

idée [Benjamini and Hochberg 1995] - control for:

$$\frac{\text{no. false positive features}}{\text{no. significant features}} = \frac{F}{F + T} = \frac{F}{S}$$

$$\text{FDR} = \text{E} \left[\frac{F}{F + T} \right] = \text{E} \left[\frac{F}{S} \right].$$

Statistical significance for genomewide studies

John D. Storey*[†] and Robert Tibshirani[‡]

*Department of Biostatistics, University of Washington, Seattle, WA 98195; and [‡]Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved May 30, 2003 (received for review January 28, 2003)

With the increase in genomewide experiments and the sequencing of multiple genomes, the analysis of large data sets has become commonplace in biology. It is often the case that thousands of features in

to the method in ref. 5 under certain assumptions. Also, ideas similar to FDRs have appeared in the genetics literature (1, 13).

Similarly to the p value, the q value gives each feature its own

We got m p values, p_1, p_2, \dots, p_m :

for a threshold t we may say that:

$$F(t) = \# \{ \text{null } p_i \leq t; i = 1, \dots, m \} \text{ and}$$

$$S(t) = \# \{ p_i \leq t; i = 1, \dots, m \}.$$

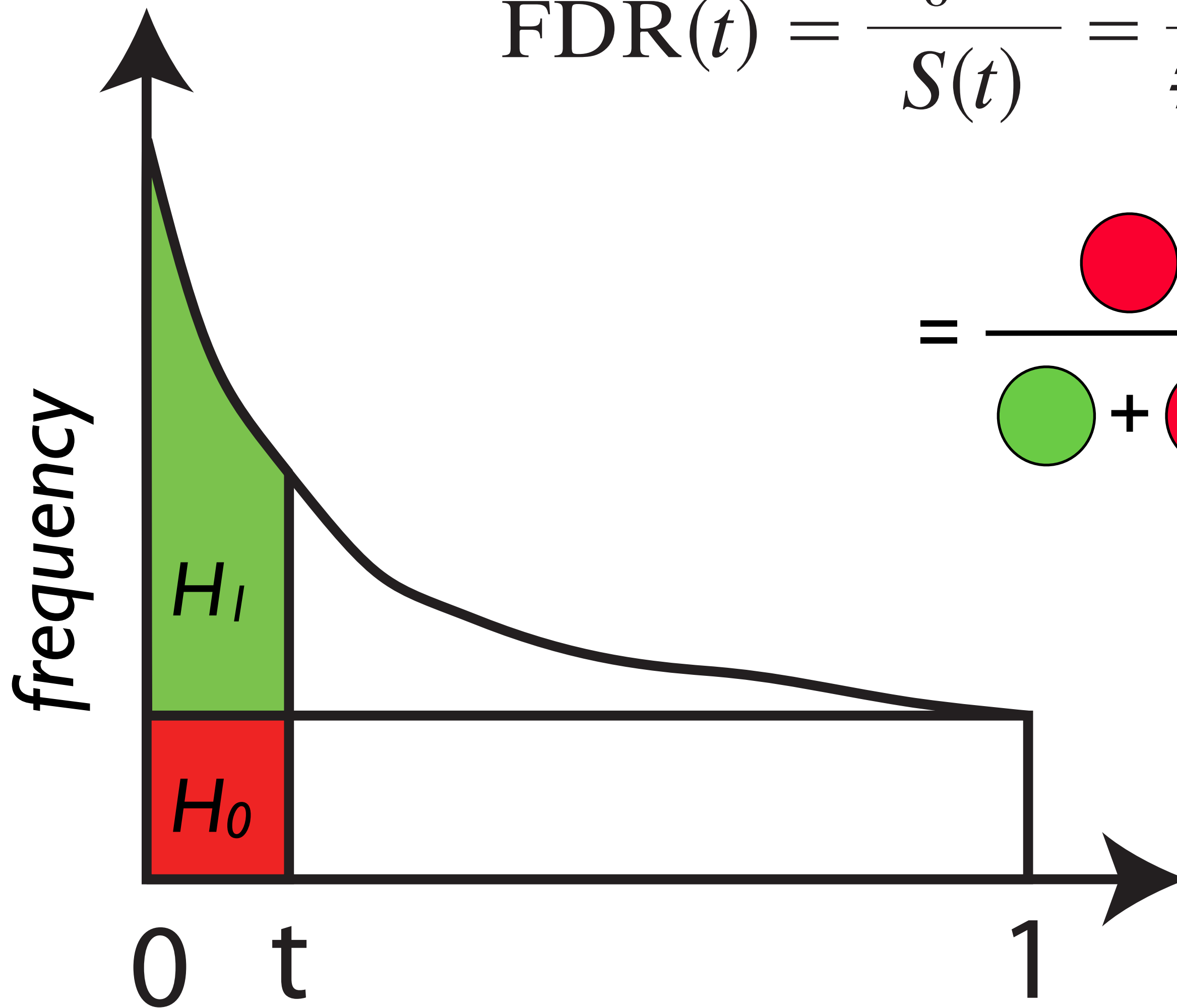
$$\text{FDR}(t) = \text{E} \left[\frac{F(t)}{S(t)} \right].$$

Evenly distributed p values: $F(t) = m \cdot t = \pi_0 m t$

$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{\# \{ p_i \leq t \}}.$$

Illustration of $\widehat{\text{FDR}}$

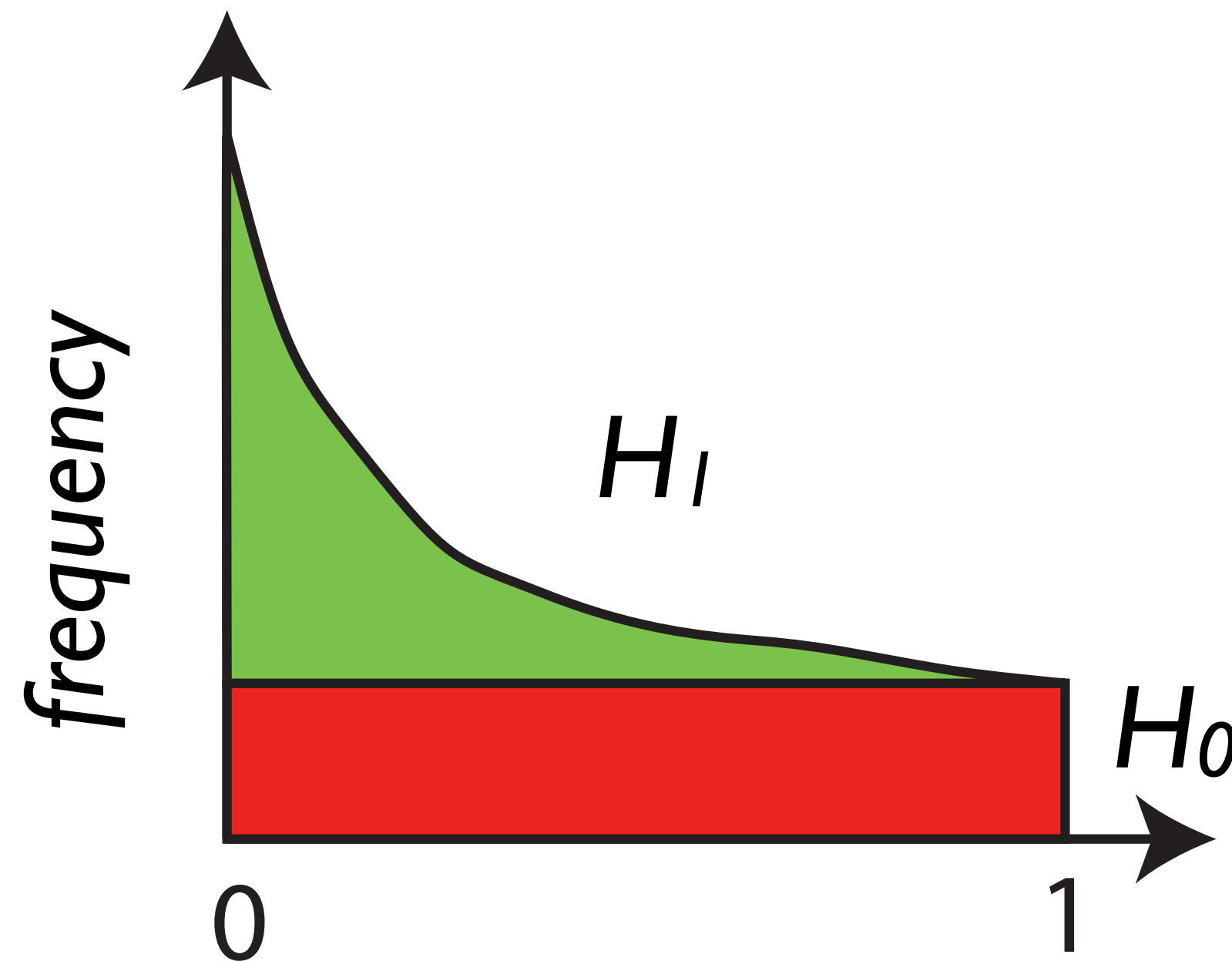
$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \leq t\}} \cdot$$



$$= \frac{\text{red circle}}{\text{green circle} + \text{red circle}}$$

Illustration of π_0

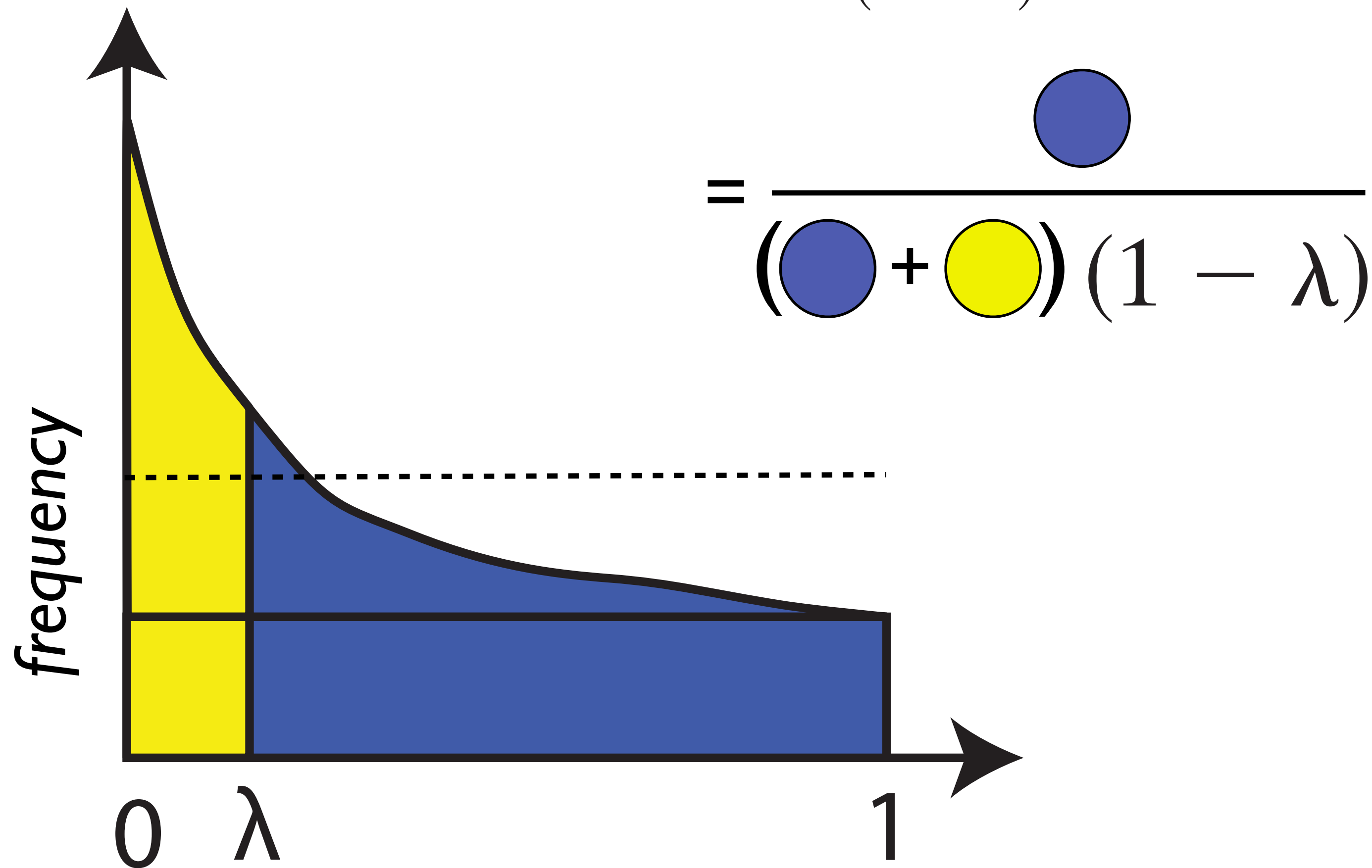
π_0 is the prior probability that a statistic is derived under H_0 i.e. $\Pr(H=H_0)$



$$\pi_0 = \frac{\text{red circle}}{\text{green circle} + \text{red circle}}$$

Π_0 estimation

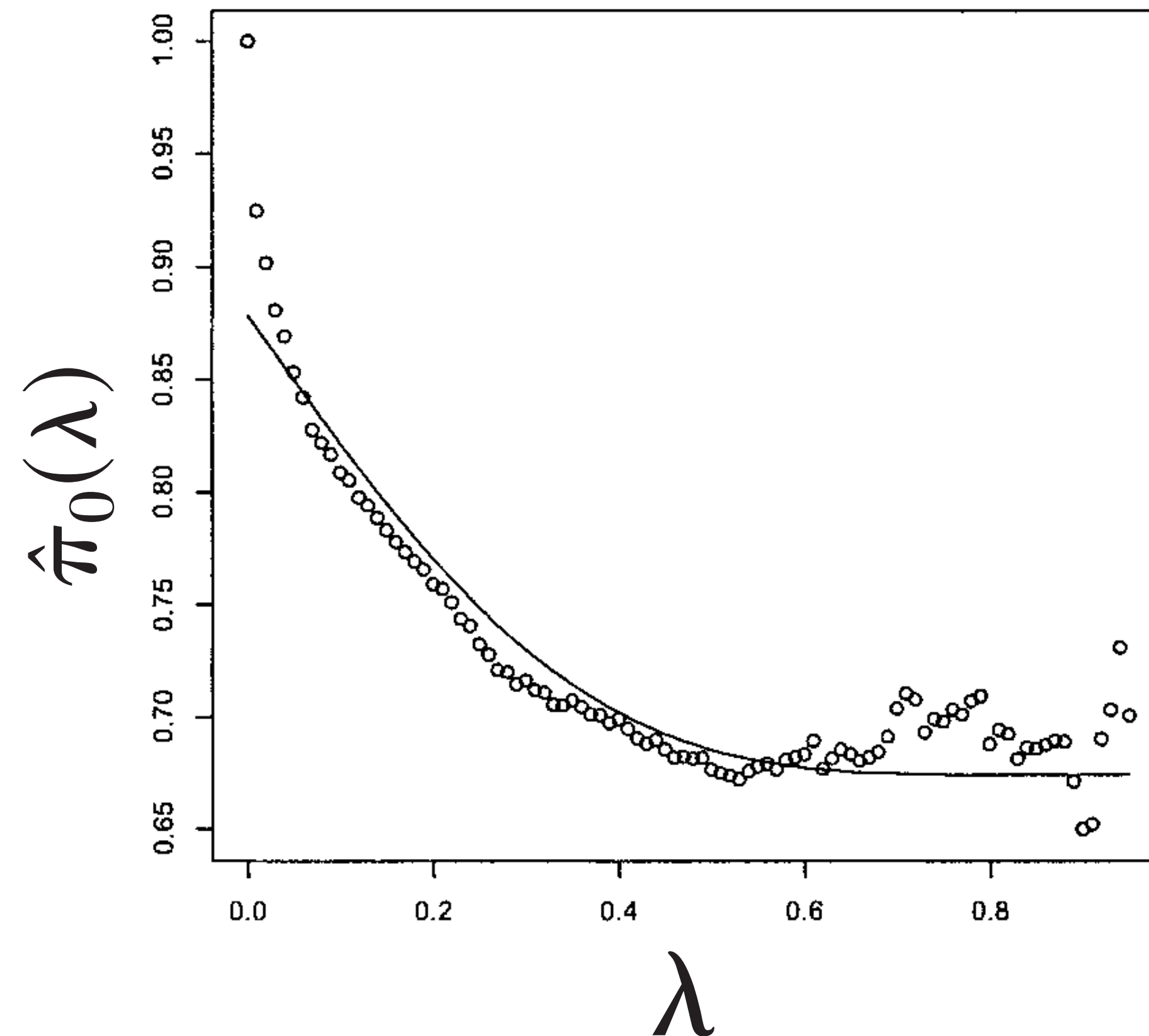
$$\hat{\pi}_0(\lambda) = \frac{\# \{p_i > \lambda; i = 1, \dots, m\}}{m(1 - \lambda)},$$



Π_0 estimation

Investigate the higher
(close to 1) p values

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \dots, m\}}{m(1 - \lambda)},$$



q value

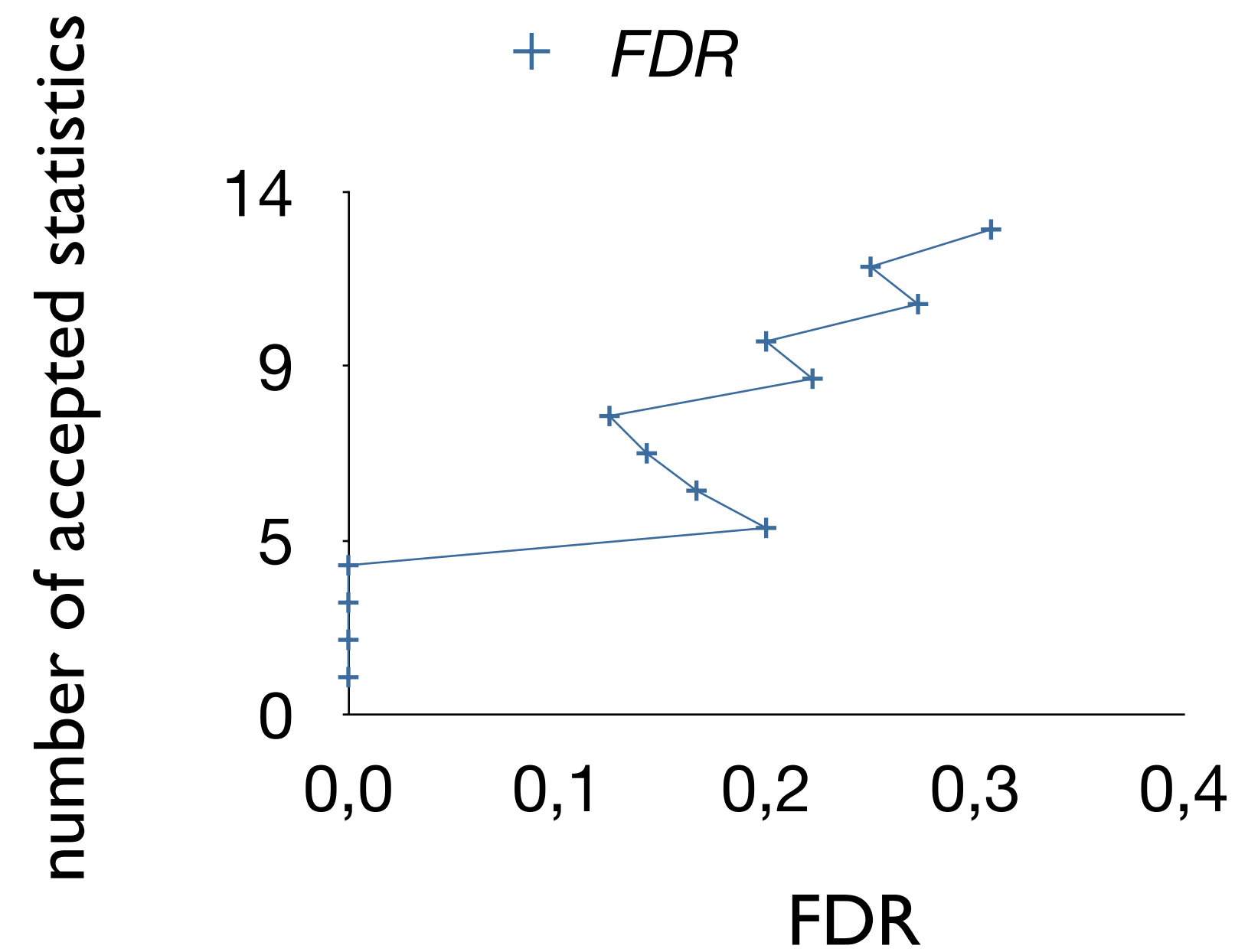
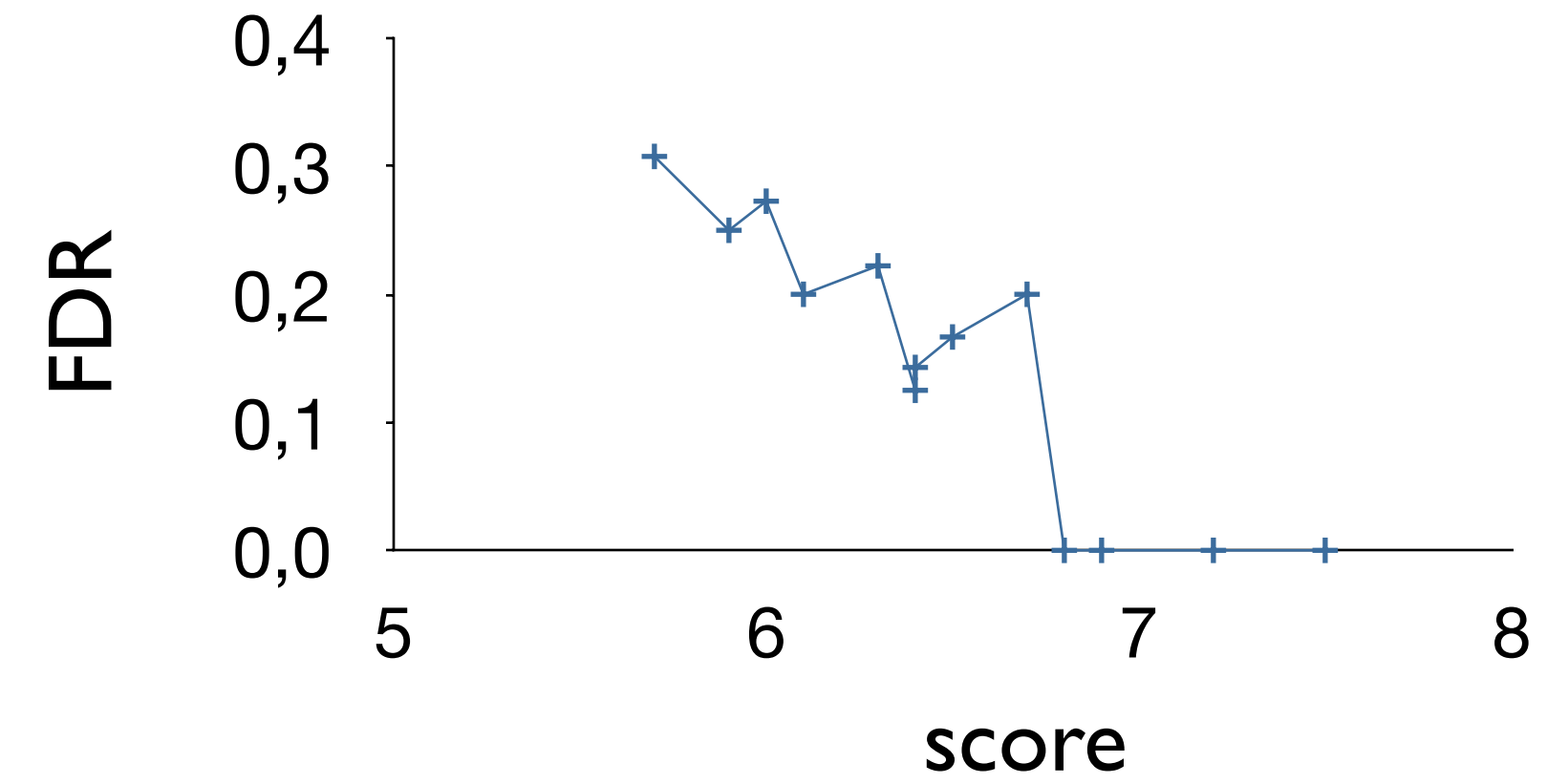
A relevant measure to individual identifications that ensures monotonically increasing function with the p value threshold.

The q value is defined as

$$\hat{q}(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t).$$

q value

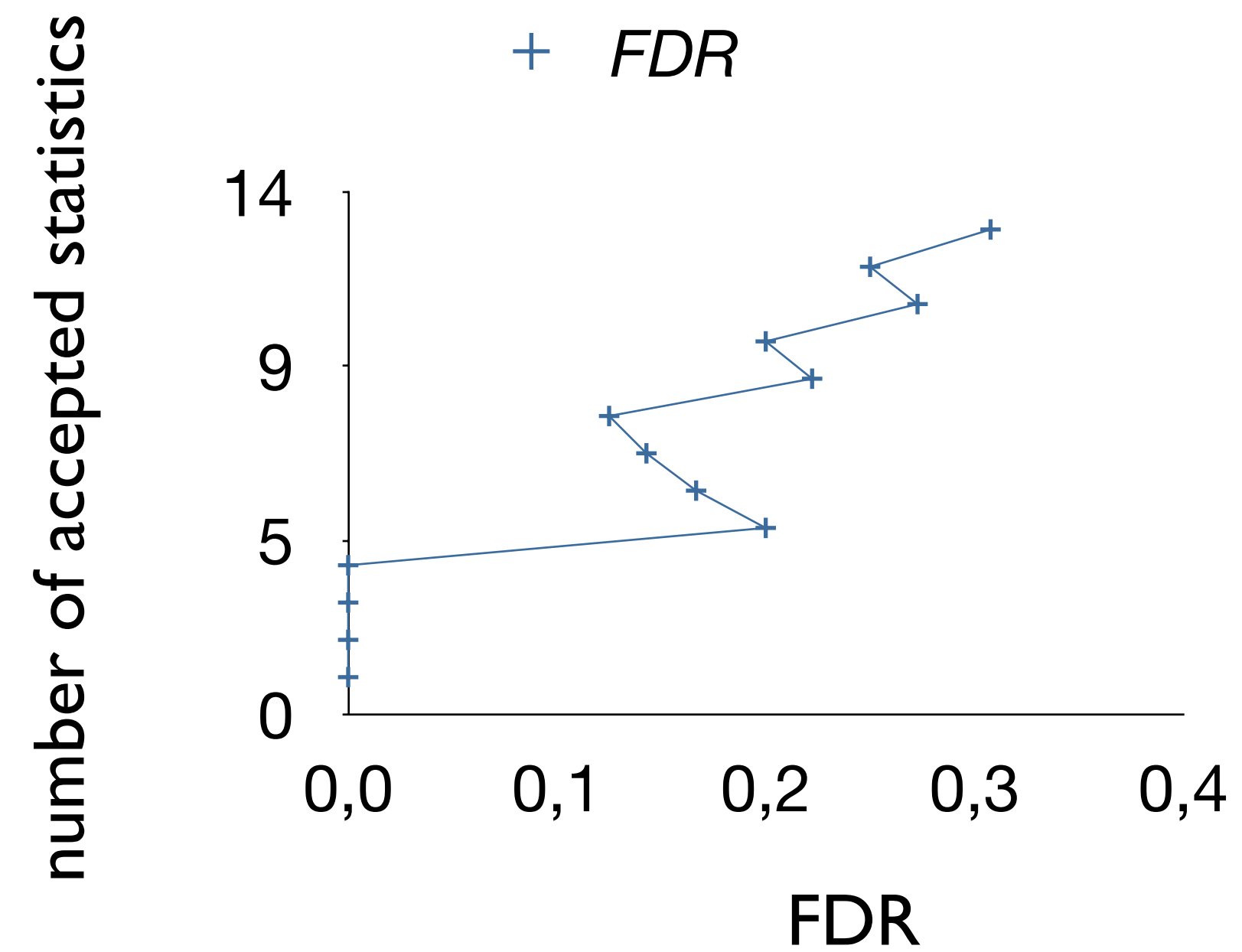
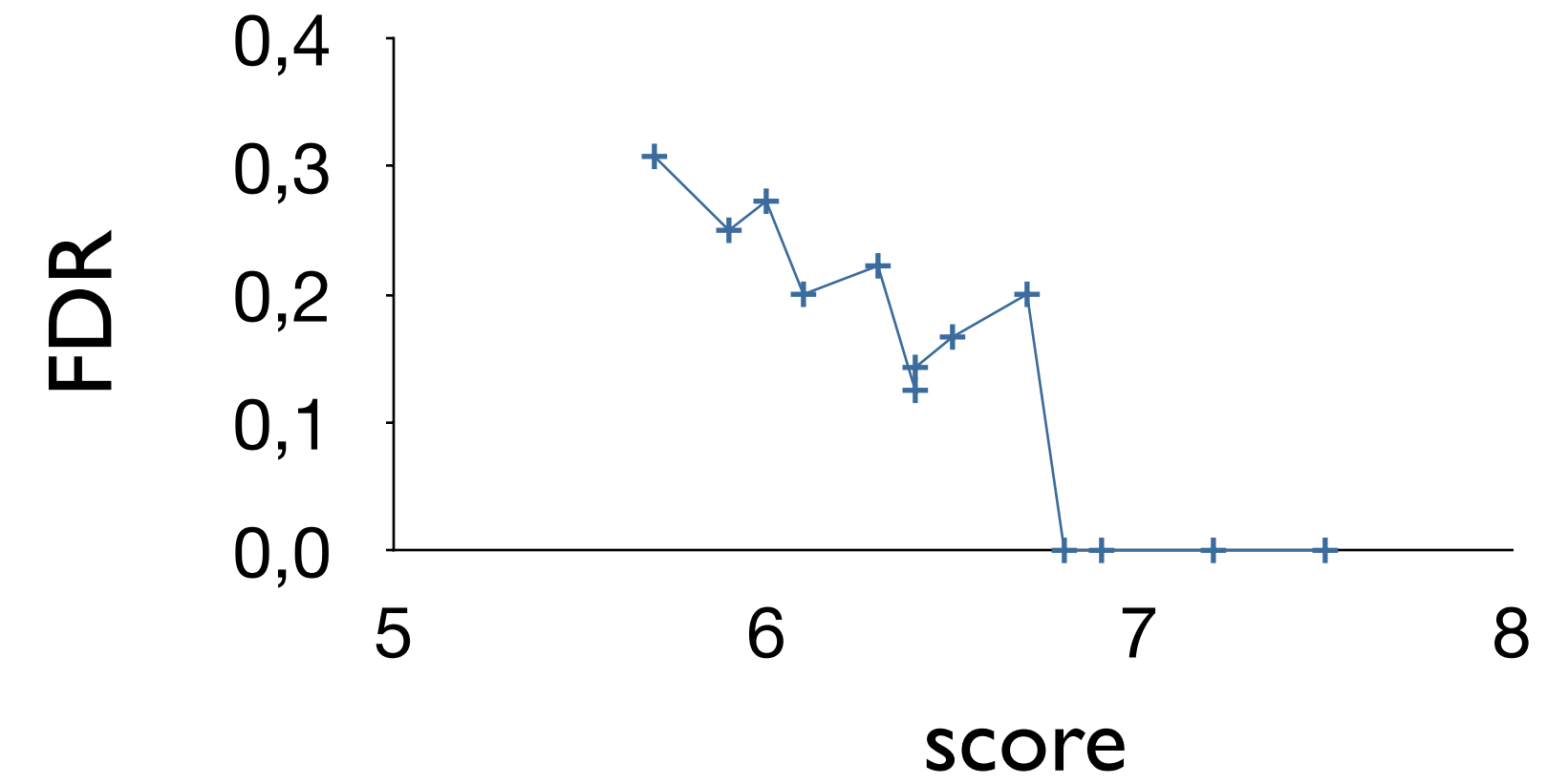
score	type
7,5	correct
7,2	correct
6,9	correct
6,8	correct
6,7	incorrect
6,5	correct
6,4	correct
6,4	correct
6,3	incorrect
6,1	correct
6	incorrect
5,9	correct
5,7	incorrect
...	...



$$q(x) = \min_{x \geq x'} \{FDR(x')\}$$

q value

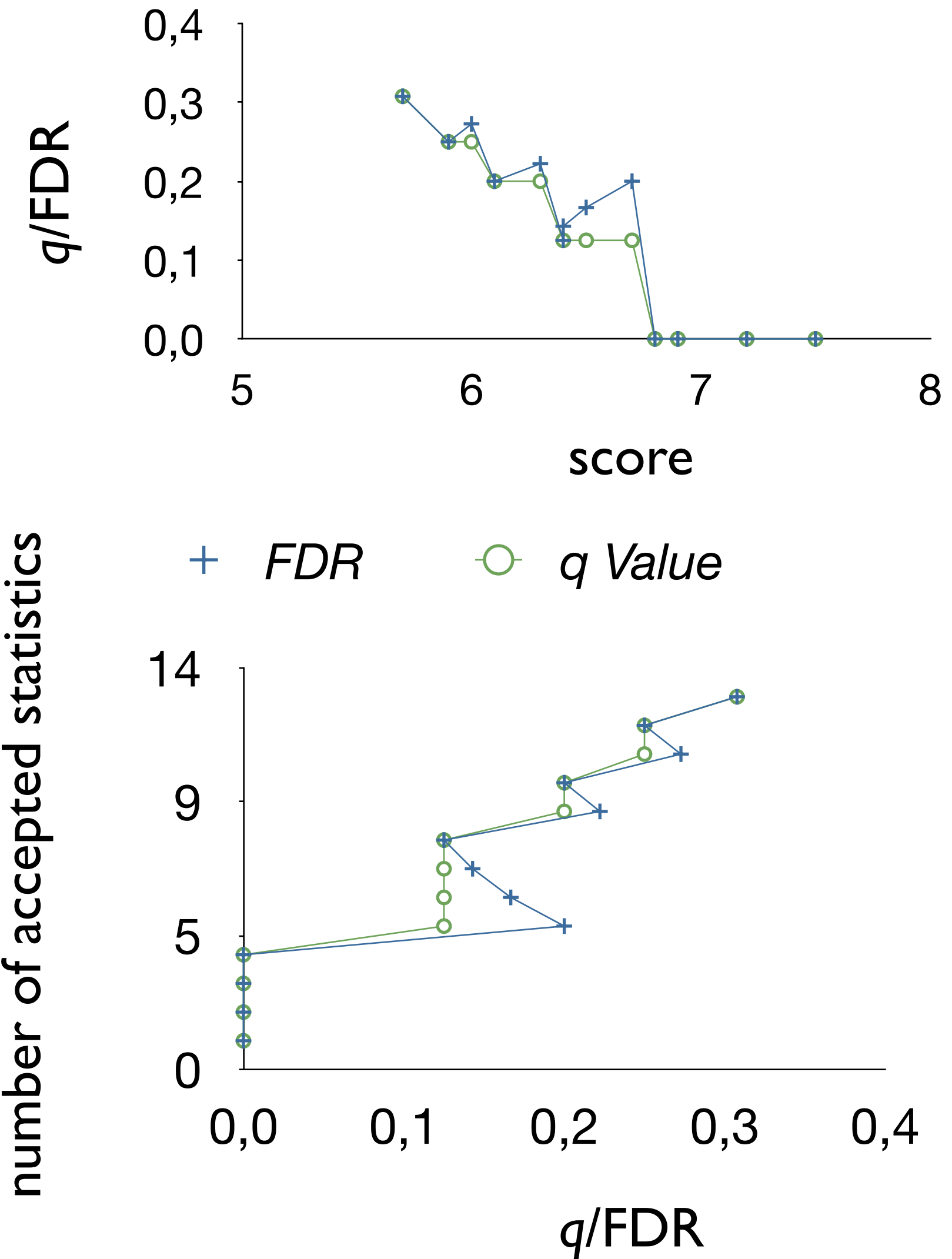
score	type
7,5	correct
7,2	correct
6,9	correct
6,8	correct
6,7	incorrect
6,5	correct
6,4	correct
6,4	correct
6,3	incorrect
6,1	correct
6	incorrect
5,9	correct
5,7	incorrect
...	...



$$q(x) = \min_{x \geq x'} \{FDR(x')\}$$

q value

score	type
7,5	correct
7,2	correct
6,9	correct
6,8	correct
6,7	incorrect
6,5	correct
6,4	correct
6,4	correct
6,3	incorrect
6,1	correct
6	incorrect
5,9	correct
5,7	incorrect
...	...



Multiple measurements per sampled individual

