Benjamini-Hochberg method (1995) for "false discovery rate",
http://www.jstor.org/stable/2346101 ;
implemented in Statistical Analysis of Microarrays (SAM) and BRBTOOLS from the National Cancer Institute's Biometric Research Branch.

Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

B-H is a classical frequentist technique. False discovery rate could be defined as

$$\text{FDR} = Q_e = E(\mathbf{Q}) = E\{\mathbf{V}/(\mathbf{V} + \mathbf{S})\} = E(\mathbf{V}/\mathbf{R}).$$

Here R = # declared significant. What if can equal 0? The Q_e will be infinite. And when all null hypotheses are true, then all discoveries are false and FDR=1. Instead B&H define

$$\text{FDR} = P(\mathbf{R} > 0)\, E(\mathbf{V}/\mathbf{R} \mid \mathbf{R} > 0).$$

The BH procedure is:

$$\text{let } k \text{ be the largest } i \text{ for which } P_{(i)} \leqslant \frac{i}{m} q^*;$$

$$\text{then reject all } H_{(i)},\ i = 1, 2, \ldots, k.$$

Then FDR is no bigger than q^*.

Example: An RCT of rt-PA versus APSACj when myocardial infarction occurs. There are 15 different clinical endpoints. The 15 P values, ranked, and the critical values with q^*=0.05, are

Pvalue	**0.0001**	**0.0004**	**0.0019**	**0.0095**	0.0201	0.0278	0.0298	...	1
cutoff	**0.0033**	**0.0067**	**0.0100**	**0.0133**	0.0167	0.0200	0.0233	...	0.05

The first 4 hypotheses are “significant” with this rule.

This method does not really help with the problems listed before. Suppose we tweak the results:

Pvalue	**0.0001**	**0.0004**	**0.0019**	**0.0095**	**0.0151**	0.0278	0.0298	...	1
cutoff	**0.0033**	**0.0067**	**0.0100**	**0.0133**	**0.0167**	0.0200	0.0233	...	0.05

OK, the 5th is significant. Now tweak the 4th. It's no longer significant.

Pvalue	**0.0001**	**0.0004**	**0.0019**	0.0135	0.0151	0.0278	0.0298	...	1
cutoff	**0.0033**	**0.0067**	**0.0100**	0.0133	0.0167	0.0200	0.0233	...	0.05

Suddenly the 5th is ALSO not significant. Why not? Does this make sense? The data for the 5th test has not changed at all. That’s a typical result for frequentist approaches to multiple comparisons. It’s unsatisfying to many people.