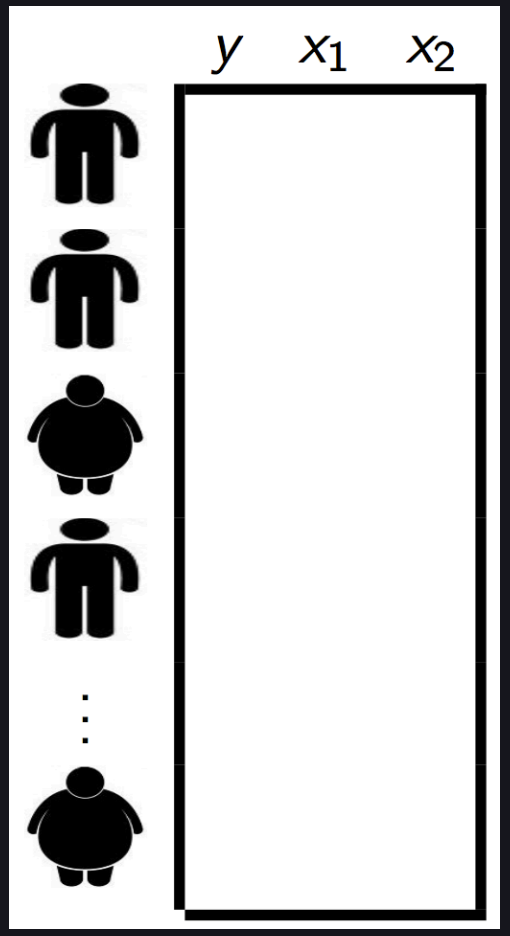
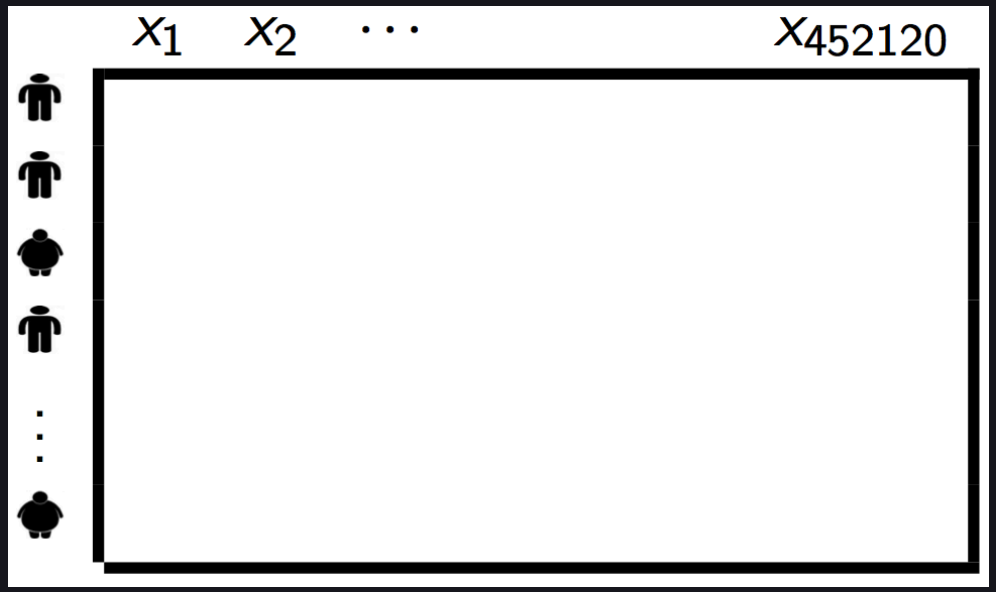
Multiple testing and resampling

Claus Thorn Ekstrøm
UCPH Biostatistics
ekstrom@sund.ku.dk



Data sizes. The $N \ll P$ problem

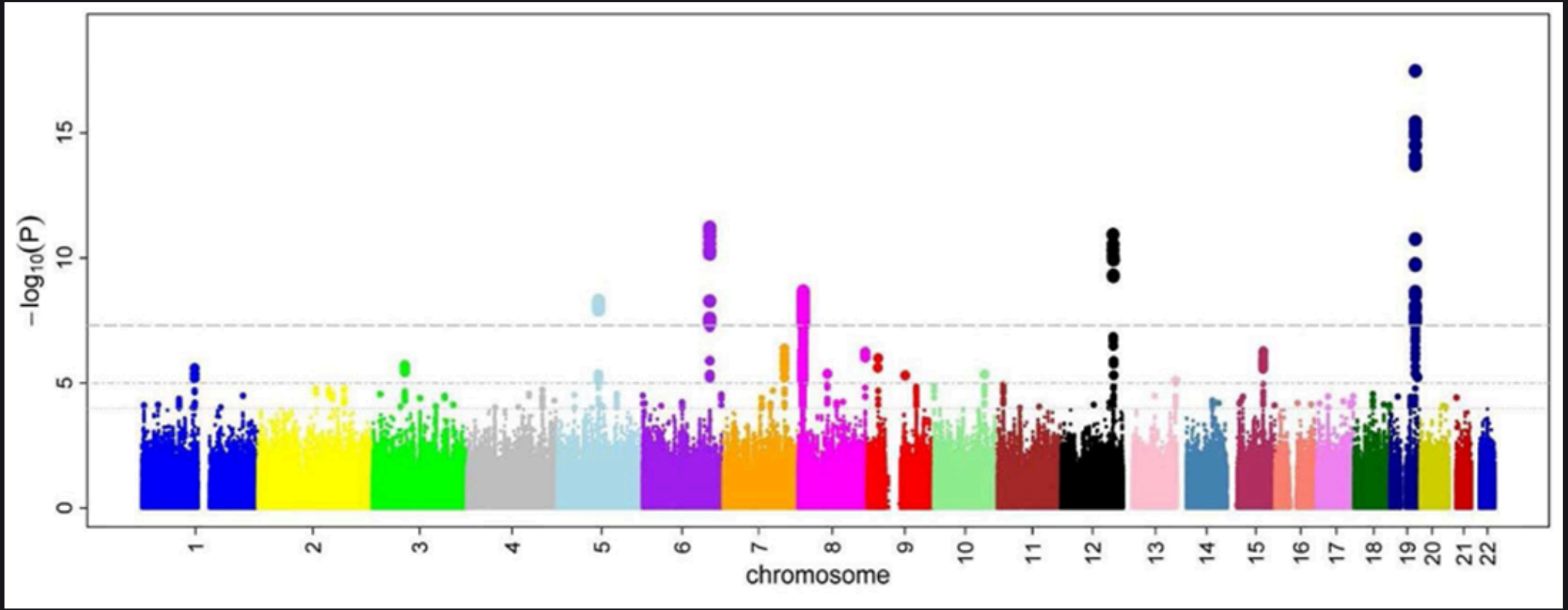




The "Big Data" revolution

1. "Big P small N" problem with many modern large-scale-datasets: registers, images, *-omics, ...
2. Need to reduce the dimension in some way
3. How do we evaluate significance when we have used the data for feature selection?
4. Multiple testing becomes an issue --- not just for high-dimensional data

Manhattan plot



Multiple comparison problems

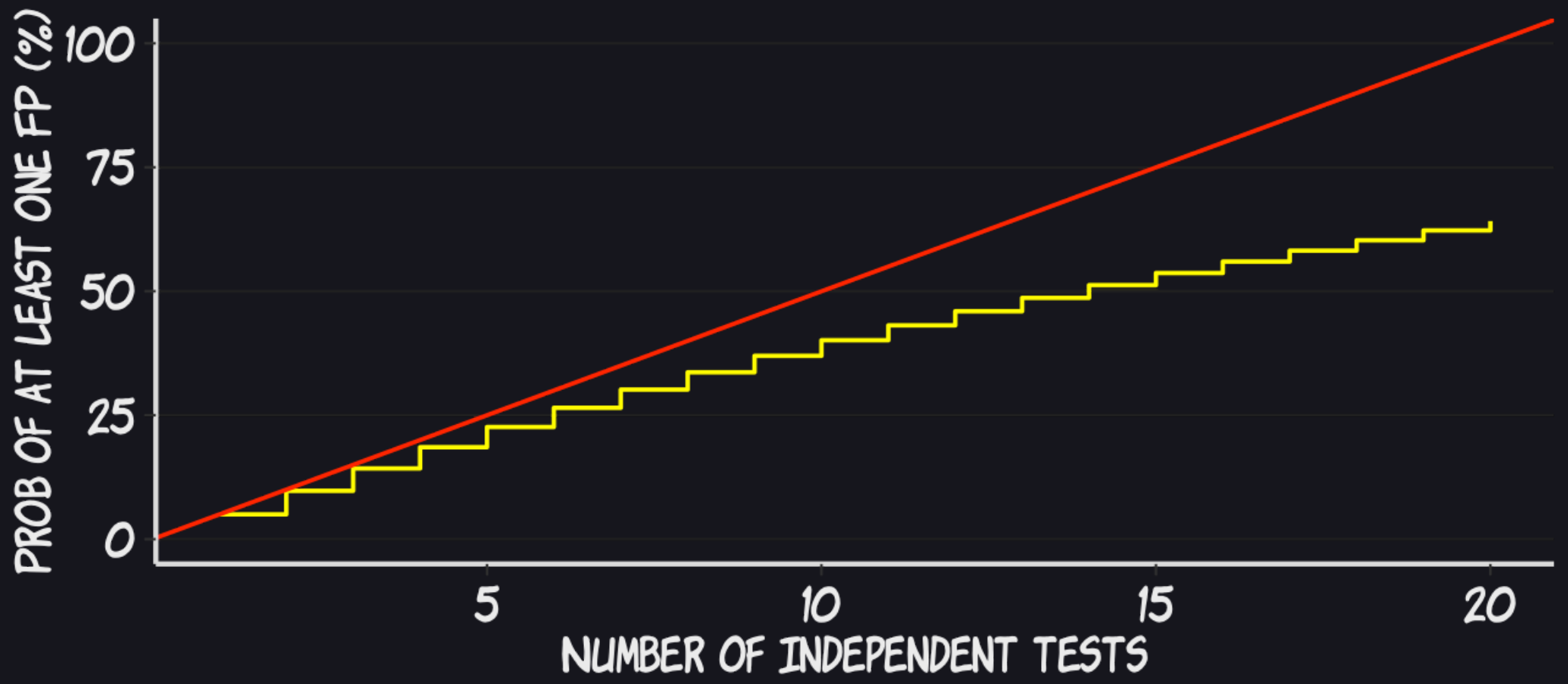
Errors committed when testing a single null hypotheses, H_0

Analysis result	**H_0 true**	**H_0 false**
Reject	α	1-β
Don't reject	1-α	β

α is the significance level

$1 - \beta$ is the power

Multiple testing



Multiple comparison problems

The family-wise error rate (FWER) is the probability of making at least one type I error (false positive).

For m tests where all null hypotheses are true we have

$$FWER = P\Big(\bigcup_{i=1}^{m}\{p_i \leq \alpha\}\Big) = 1 - P(\text{no false positives}) = 1 - (1-\alpha)^m \leq m\alpha$$

where the second equality only holds under independence, but $FWER \leq m\alpha$ always holds due to Boole's inequality.

Multiple comparison problems

Number of errors committed when testing m null hypotheses.

Analysis result	H_0 true	H_0 false	Total
Reject	V	S	R
Don't reject	U	T	m-R
Total	m_0	$m - m_0$	m

Here R, the number of rejected hypotheses/discoveries. V, S, U and T are unobserved. The FWER is

$$FWER = P(V > 0) = 1 - P(V = 0)$$

Bonferroni correction

The most conservative method but is free of dependence and distributional assumptions. Set the significance level for each individual test at α/m. By Boole's inequality (no independence needed)

$$FWER = P\Big(\bigcup_{i \in H_0} \{p_i \leq \alpha/m\}\Big) \leq \sum_{i \in H_0} P(p_i \leq \alpha/m) = m_0 \frac{\alpha}{m} \leq \alpha$$

In other words we reject the ith hypothesis if

$$mp_i \leq \alpha \Leftrightarrow p_i \leq \frac{\alpha}{m}$$

Sidak correction

$$1 - (1 - \alpha)^m = \alpha^* \Leftrightarrow \alpha = 1 - \sqrt[m]{1 - \alpha^*}$$

Slightly less conservative than Bonferroni (but not much). Requires independence (or positive dependence) of the tests.

Holm correction

1. Compute and order the individual p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. Find $\hat{k} = \min\{k : p_{(k)} > \frac{\alpha}{m+1-k}\}$
3. If $\hat{k}$ exists then reject hypotheses corresponding to $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(\hat{k}-1)}$. Otherwise reject all m hypotheses.

Holm correction

Controls the FWER: Assume the (ordered) k is the first wrongly rejected true hypothesis. Then $k \leq m - (m_0 - 1)$. Hypothesis k was rejected so

$$p_{(k)} \leq \frac{\alpha}{m + 1 - k} \leq \frac{\alpha}{m + 1 - (m - (m_0 - 1))} \leq \frac{\alpha}{m_0}$$

Since there are m_0 true hypotheses then (Bonferroni argument) the probability that one is significant is at most α so FWER is controlled.

Takeaway: Holm always rejects at least as many hypotheses as Bonferroni and still controls the FWER without assumptions. There is no reason to use plain Bonferroni.

Practical problems

- While guarantee of FWER-control is appealing, the resulting thresholds often suffer from low power.

  In practice, this tends to wipe out evidence of the most interesting effects

- FDR control offers a way to increase power while maintaining some principled bound on error

False discovery rate

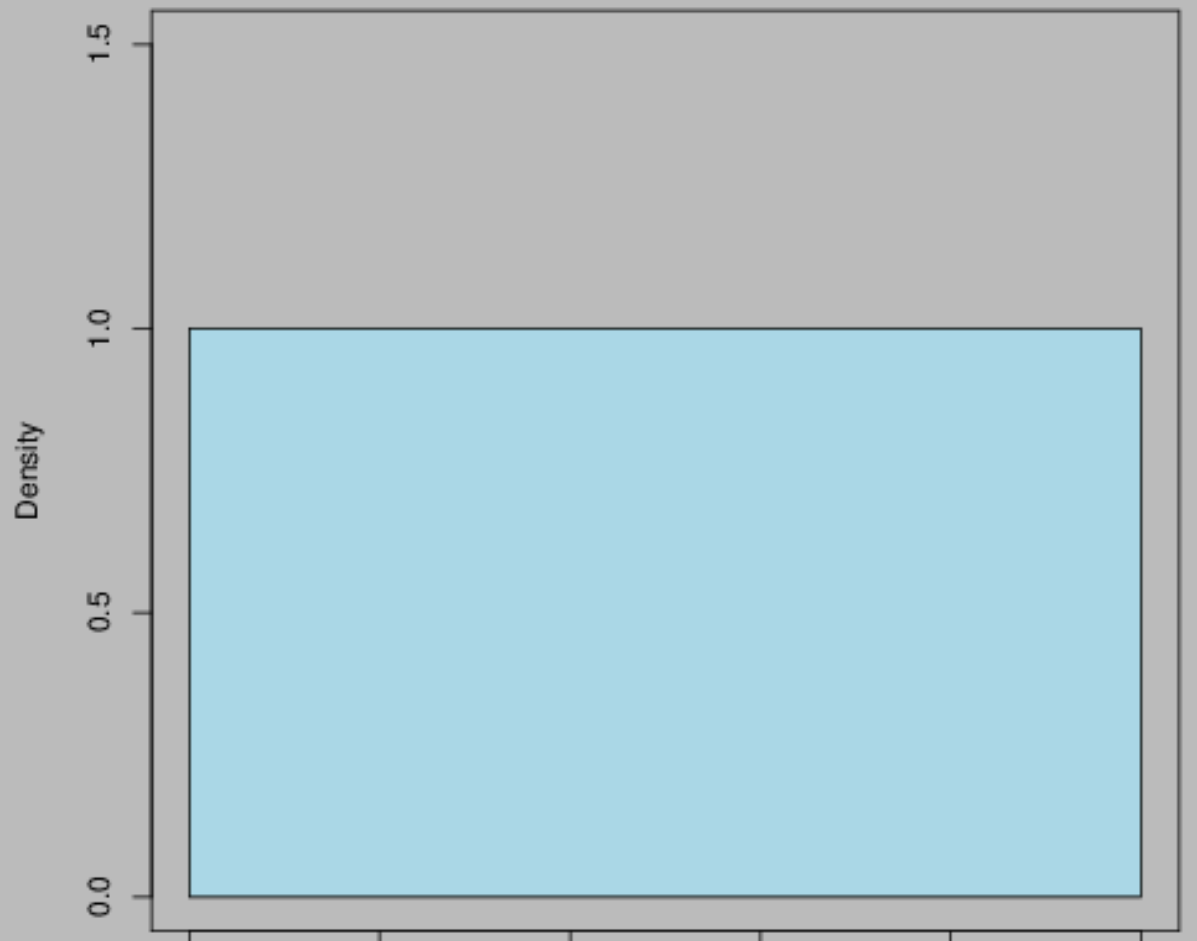
Number of errors committed when testing m null hypotheses.

Analysis result	H_0 true	H_0 false	Total
Reject	V	S	R
Don't reject	U	T	m-R
Total	m_0	$m - m_0$	m

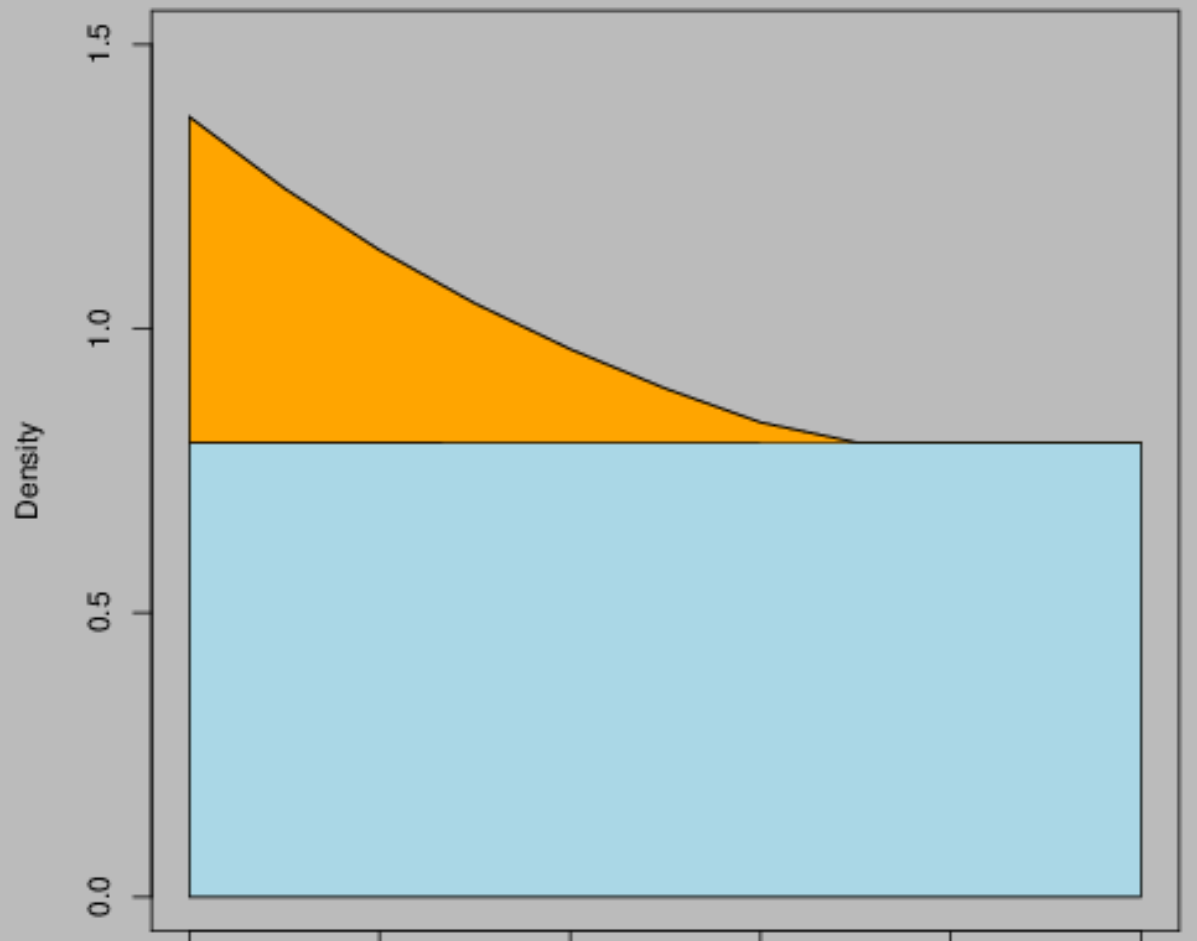
Proportion of false discoveries is $Q = \frac{V}{R}$. [Set to 0 for $R = 0$]

The false discovery rate is $FDR = E(Q) = E(\frac{V}{R})$

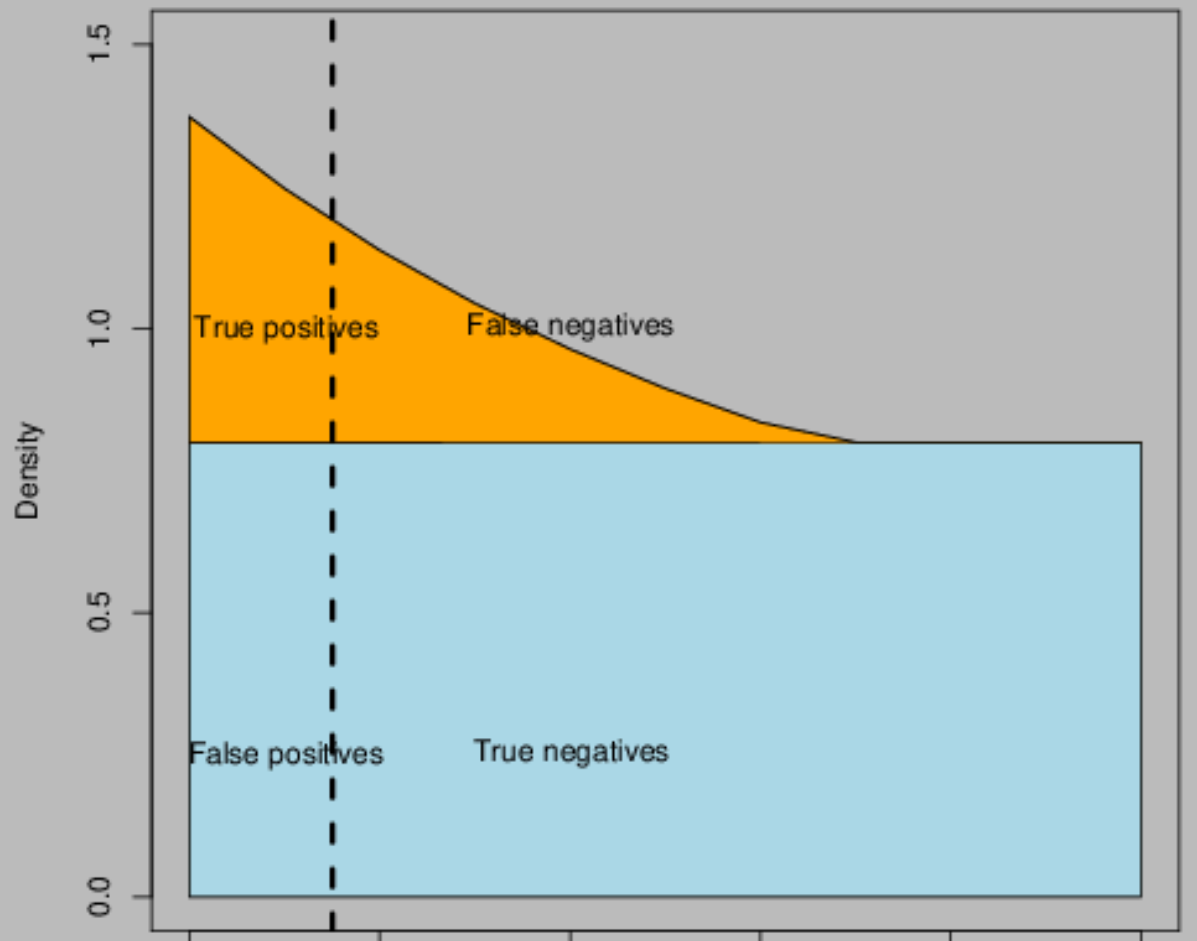
Estimating FDR



Estimating FDR



Estimating FDR



Controlling FDR — BH step-up

Benjamini-Hochberg step-up procedure to control the FDR at α.

1. Compute and order the individual p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. Find $\hat{k} = \max\{k : \frac{m}{k} \cdot p_{(k)} \leq \alpha\}$
3. If $\hat{k}$ exists then reject hypotheses corresponding to $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(\hat{k})}$

Controlling FDR — BH step-up

p-values

$$
\begin{aligned}
\tilde{p}_{(1)} &= \min\{\tilde{p}_{(2)}, m p_{(1)}\} \\
&\vdots \qquad\qquad \vdots \\
\tilde{p}_{(m-1)} &= \min\{\tilde{p}_{(m)}, \tfrac{m}{m-1} p_{(m-1)}\} \\
\tilde{p}_{(m)} &= p_{(m)}
\end{aligned}
$$

Note that BH rejects at least as many hypotheses as Holm, so BH does **not** control the FWER (only when all m null hypotheses are true).

Controlling FDR — BH step-up

Assume independent tests and uniform p-values for the m_0 true null hypotheses. Control FDR at level q.

Let $R^{(i)}$ be the number of rejections when p_i is set to 0. It depends only on the other p-values, and if $p_i \le qr/m$ then $R = r \Leftrightarrow R^{(i)} = r$. So

$$E\Big(\frac{V}{R}\Big) = \sum_{i\in H_0}\sum_{r=1}^{m} \frac{1}{r} P\Big(p_i \le \frac{qr}{m}, R = r\Big) = \sum_{i\in H_0}\sum_{r=1}^{m} \frac{1}{r} P\Big(p_i \le \frac{qr}{m}\Big) P(R^{(i)} = r)$$

$$= \sum_{i\in H_0}\sum_{r=1}^{m} \frac{1}{r}\frac{qr}{m} P(R^{(i)} = r) = \sum_{i\in H_0} \frac{q}{m} = q\frac{m_0}{m} \le q$$

q values

The q-value is defined to be the FDR analogue of the p-value.

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

The q-value of an individual hypothesis test is the minimum FDR at which the test may be called significant.

q values

- When all m null hypotheses are true then FDR control is equivalent to FWER control.
- BH always rejects at least as many hypotheses as Bonferroni/Holm, and never more than uncorrected testing at level α. The gain over FWER control is large when there are many true effects and negligible when signals are sparse.
- The FDR bound holds for certain classes of dependent tests. In practice, it is quite hard to "break"

Exercises

Bootstrapping: evaluating complex methods

When we have complex data (or perhaps just big data combined with simple methods) and non-parametric methods then we still wish to evaluate them.

How stable are the results?

If we have information about the distribution under the null:

- Simulate data, run algorithm to get an idea about how it behaves

If we *don't* have information about the distribution under the null