<u>Multiple comparisons</u>

"The more you look, the more you discover ... that's actually false!"

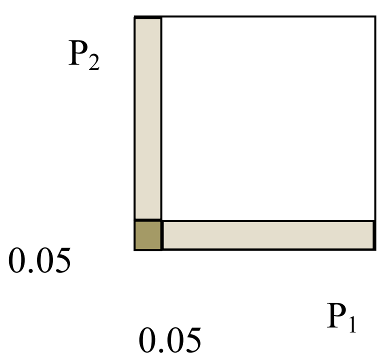
Data dredging, data snooping, fishing expedition, P-hacking, "torturing the data for a false confession".

Suppose you do $\boldsymbol{K}$ hypothesis tests. Let's evaluate the strategy:

"report the best P-value".

Suppose all the H0's are true! (No "discoveries" to be made.)
Suppose, for now, that the test statistics are all statistically independent.
(Assume continuous data & continuous P-values.)

Then the P-values, $P_1,\ldots,P_{\boldsymbol{K}}$, are i.i.d. Uniform(0,1). What is the actual distribution of the reported ***<u>best</u>*** P-value? Hint: it's not Uniform(0,1).
Remember the formula for Probability of a union (N01a).



Pr(at least one P-value $\le$ 0.05)
$= \Pr(P_{min} \le 0.05)$
$= \Pr(P_1 \le 0.05 \text{ or } P_2 \le 0.05)$
$= 2 \times 0.05 - 0.05^2$
$= 0.0975$

The classical view of multiple testing (independent case)

If you test H0 versus H_{Ai} for i=1,…,k,
and then you report the best of the k P-values ("nominal"),

$$P_{min} = \min\left\{P_i : i=1,\ldots,k\right\} = P_{i_{best}}, \text{ where } i_{best} = \arg\min\left\{P_i : i=1,\ldots,k\right\}),$$

then for this procedure

a) (pre-data) the true Type I error is bigger than the nominal Type I error,
b) (post-data) the true P-value is bigger than the nominal P-value.

For (a), the reason is that the actual Type I error = Pr(rejection region | H0), and

$$\text{overall rejection region} = \bigcup_{i=1}^{k} \text{rejection region } \boldsymbol{i}.$$

For (b), the reason is that the true "tail of surprise"
includes tails for ***all other*** hypotheses,
not just the tail for $H_{i_{best}}$.

(Multiple testing is very broad; includes for example taking interim looks at the data, which is frequent in clinical trials due to ethics.)