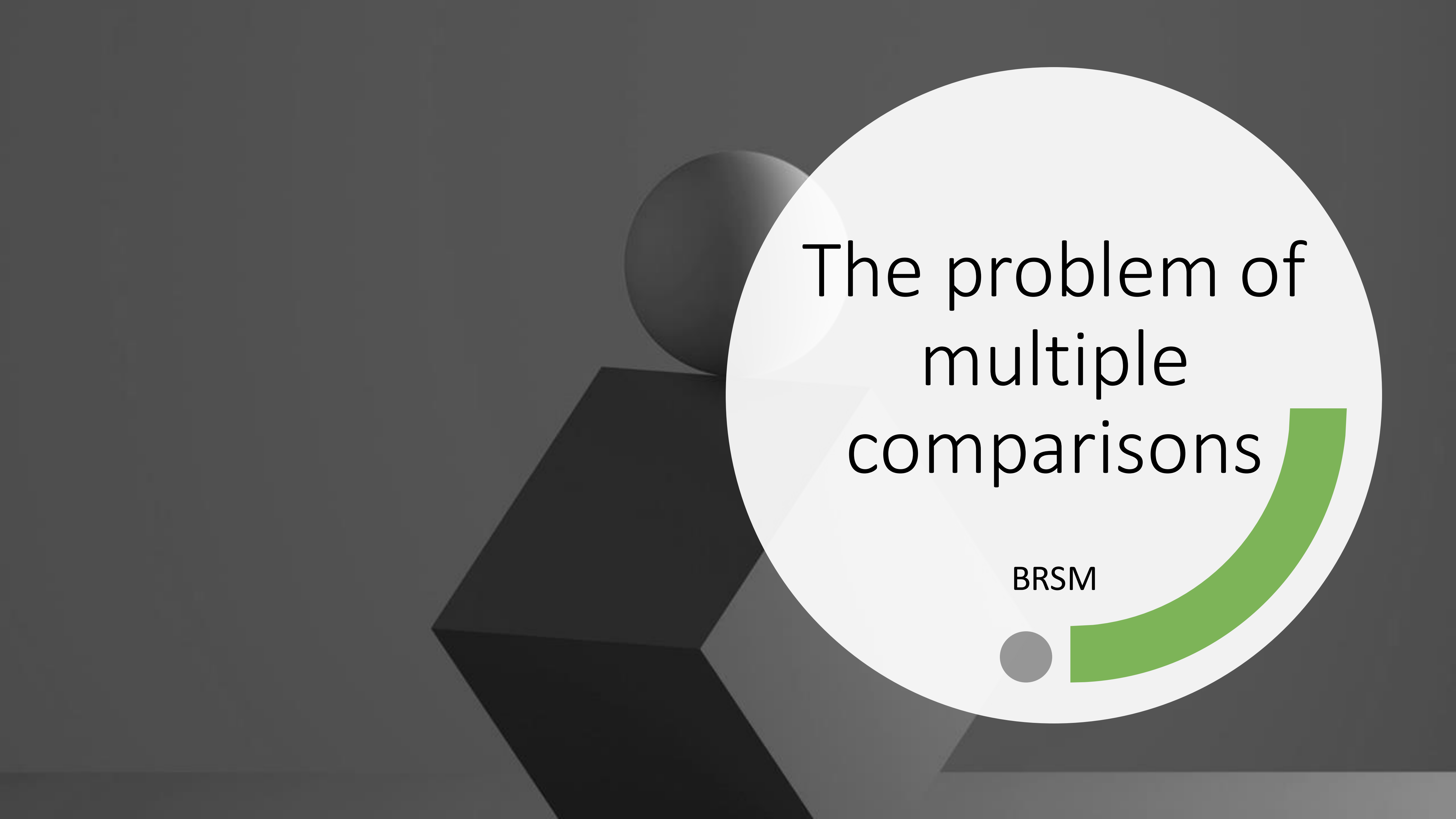


The problem of multiple comparisons

BRSM

A drug for memory improvement

- CEO: I think this drug can improve memory
- Scientists: oops, $p>0.05$
- CEO: Hmm.. Not to despair. Reanalyze the data and see if it improves concentration
- Scientists: no luck, $p > 0.05$
- CEO: Ok, here's a list of different things to try, 20 of them to be exact.
- Scientists: yay, one of them is $p < 0.05$!! It seems to improve executive control.
- CEO: See? I told you, now we go raise 100 crores. We call it the "miracle executive control drug"!

In the earlier example, alpha = 0.05

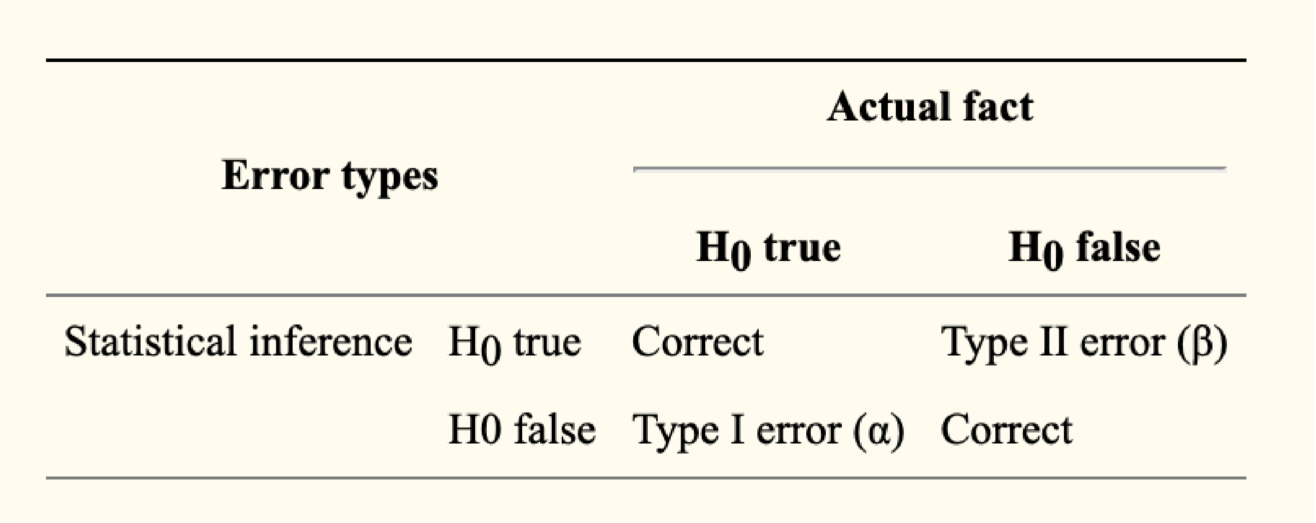
- 20 different comparisons made with the same data
- 1 of them yielded $p < 0.05$
- What is the problem with this?

Another example

- We have a coin, need to test if it is fair
- We toss it 10 times. If we get 9 heads and 1 tail, we might think that it is an unfair coin as the chances of getting it are very low for a fair coin.
- Now, clone that coin 19 times.
- Toss each one of those 20 coins 10 times each. If you get 9 heads and 1 tail for one of them, will you be confident that the coin is unfair?
- No, because with 20 coins, there is a much higher chance of obtaining 9H + 1T for a fair coin by random chance.
- This is the same issue that was present in the drug company example earlier.

Type I and type II errors

- Type I: When H0 is rejected even though it is true (a false positive from the perspective of H1)
- Type II: When H0 is accepted even though it is false (a false negative from the perspective of H1)



Error types		Actual fact	
		H_0 true	H_0 false
Statistical inference	H_0 true	Correct	Type II error (β)
	H0 false	Type I error (α)	Correct

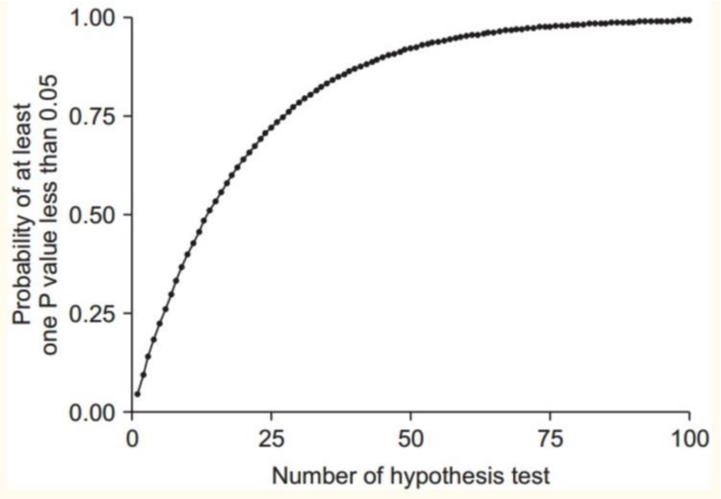
Alpha is a criterion set by us to say that this is the P(false positive) that we can live with under the null hypothesis – for one test.

The more tests you do on the data, the more likely you are to mistakenly claim an effect when there is none --> the multiple comparisons problem.

Ok, so how do you "correct" for this multiple-comparisons problem?

- Intuition: if you continue to reject the null hypothesis at alpha = 0.05, then when you have multiple tests, you are more likely to endorse effects (i.e., reject the null) when there are none (i.e., when the null is true).
- So what if you adjust the p values you obtain from all your tests such that you reduce the chance that you falsely claim an effect?
- Should you increase or decrease the p values of your multiple tests to reduce the chance of a false positive?
- Ans: increase! So if you obtain a p value of 0.04 for one of many tests you do, and if you somehow adjust that p value to 0.06, you don't falsely claim an effect.

Type 1 error rate goes up with the number of tests

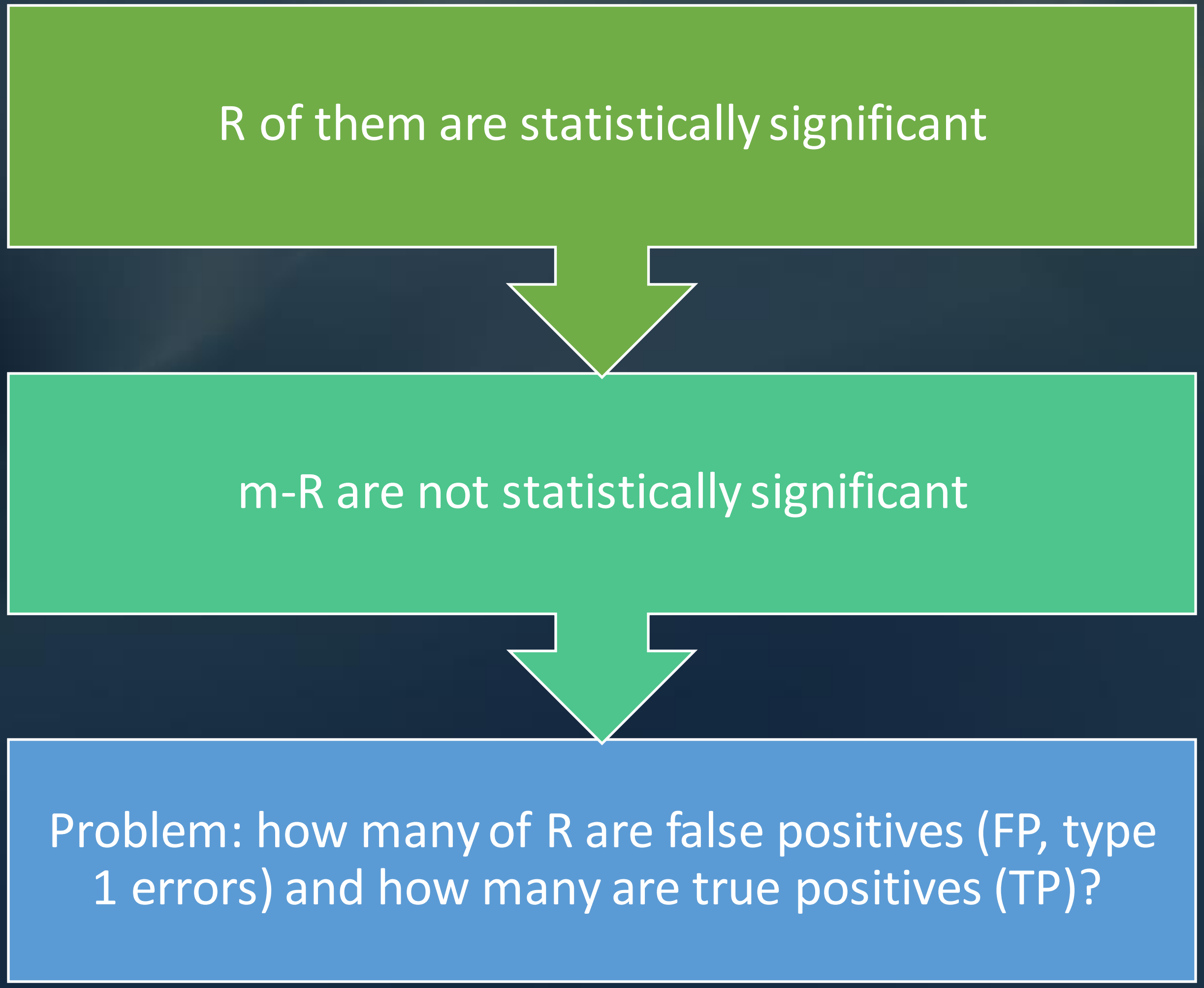


Inflated $\alpha = 1 - (1 - \alpha)^{N}$, $N =$ number of hypotheses tested

p-value adjustments

- These corrections either control the Family-Wise Error Rate (FWER) OR the False Discovery Rate (FDR)

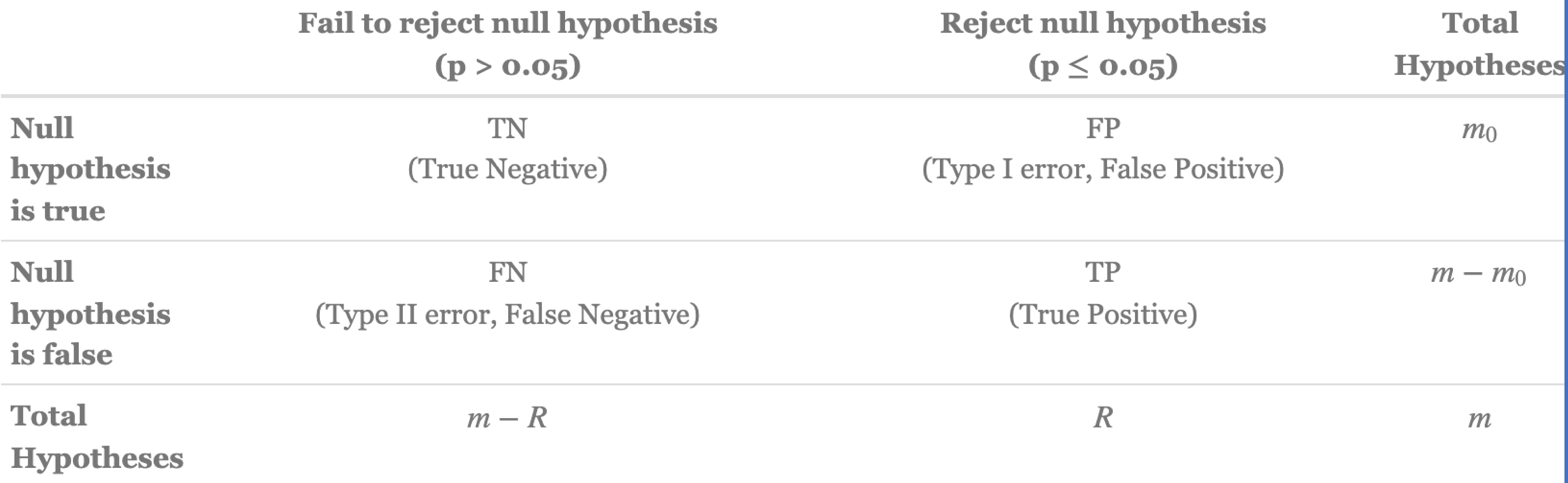
Assuming m different tests



R of them are statistically significant

m-R are not statistically significant

Problem: how many of R are false positives (FP, type 1 errors) and how many are true positives (TP)?



	Fail to reject null hypothesis ($p > 0.05$)	Reject null hypothesis ($p \leq 0.05$)	Total Hypotheses
Null hypothesis is true	TN (True Negative)	FP (Type I error, False Positive)	m_0
Null hypothesis is false	FN (Type II error, False Negative)	TP (True Positive)	$m - m_0$
Total Hypotheses	$m - R$	R	m

Error table for a family of m tests

Family-Wise Error Rate (FWER) vs False Discovery Rate (FDR)

- FWER = Probability of falsely rejecting even one null hypothesis = P(FP >= 1) across all m tests.
- FDR = Expected proportion of false discoveries among all discoveries = E[FP/R]. That is, you take all the rejected hypotheses and find how many of those "discoveries" were false. The expected proportion of this is the FDR.
- FWER = FDR **if all null hypotheses are true** and so are connected to each other but they are different and the difference is subtle.

Family-Wise Error Rate (FWER)

Assume we have 3 null hypotheses, all of which are true (drug does not improve memory, does not improve concentration, does not improve executive control)

Alpha = 0.05 (our criterion for type 1 errors that we can live with)

Now, for the family of 3 tests, what is the type 1 error probability?

The prob of rejecting any given hypothesis erroneously = 0.05, prob of accepting the null = 0.95

The prob of rejecting any one of three tests erroneously = 1 – 0.95^3 = 0.142

Controlling for this FWER involves reducing this 0.142 to 0.05 as you would expect for a single test

Bonferroni correction: a simple but very conservative way to control for FWER

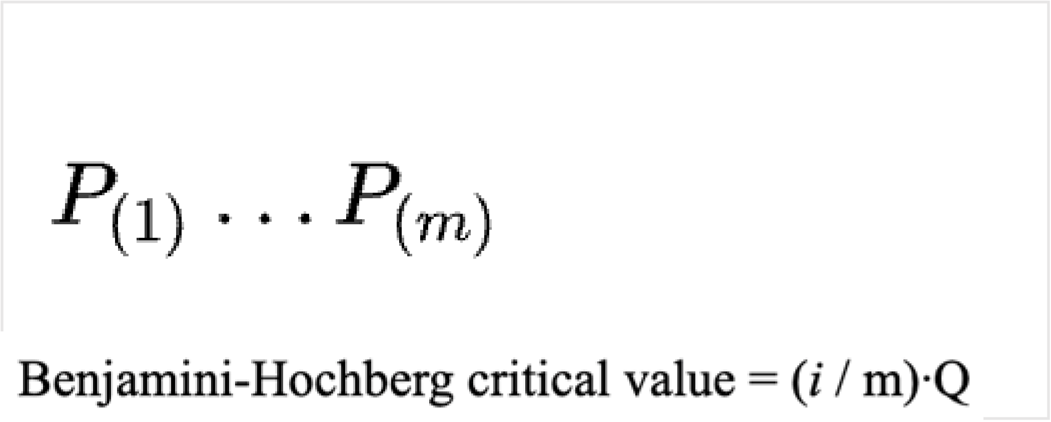
- If you have 20 tests, adjust the alpha by dividing by the total number of tests (I.e., each individual test now has to pass a more stringent criterion such that the family-wise type 1 error remains at 0.05)
- e.g. 50 t-tests. Adjusted alpha = 0.05/50 = 0.001
- So any given test of the 50 different tests should be considered significant only if $p < 0.001$ and not $p < 0.05$ so that the overall family-wise type 1 error is at 0.05

Adjusted alpha (α) = α / k (number of hypothesis tested)

The problems with Bonferroni correction

- Too stringent: The side-effect is that it also increases type 2 error (I.e., we miss out on true effects due to the stringent criterion)
- Assumes independent tests but in many cases, our tests are not independent (e.g. in neuroscience, when testing different brain regions, the regions are not independent, they have correlated activity, etc)
- The correction does not depend on the structure of the data and instead only on the number of tests. For any given true effect, you will change the ability to find that effect drastically by simply changing the number of tests you do.
- Holm correction: a sequential procedure that still controls for FWER but does not increase type 2 error as much as Bonferroni does.
- Do you care about not making ANY false positives? Then controlling for FWER is appropriate.
- In many domains, you can live with some false positives (such as in genomics), there the more appropriate quantity to control for is FDR.

Controlling for False Discovery Rate (FDR) - Benjamini-Hochberg procedure



$$P_{(1)} \dots P_{(m)}$$

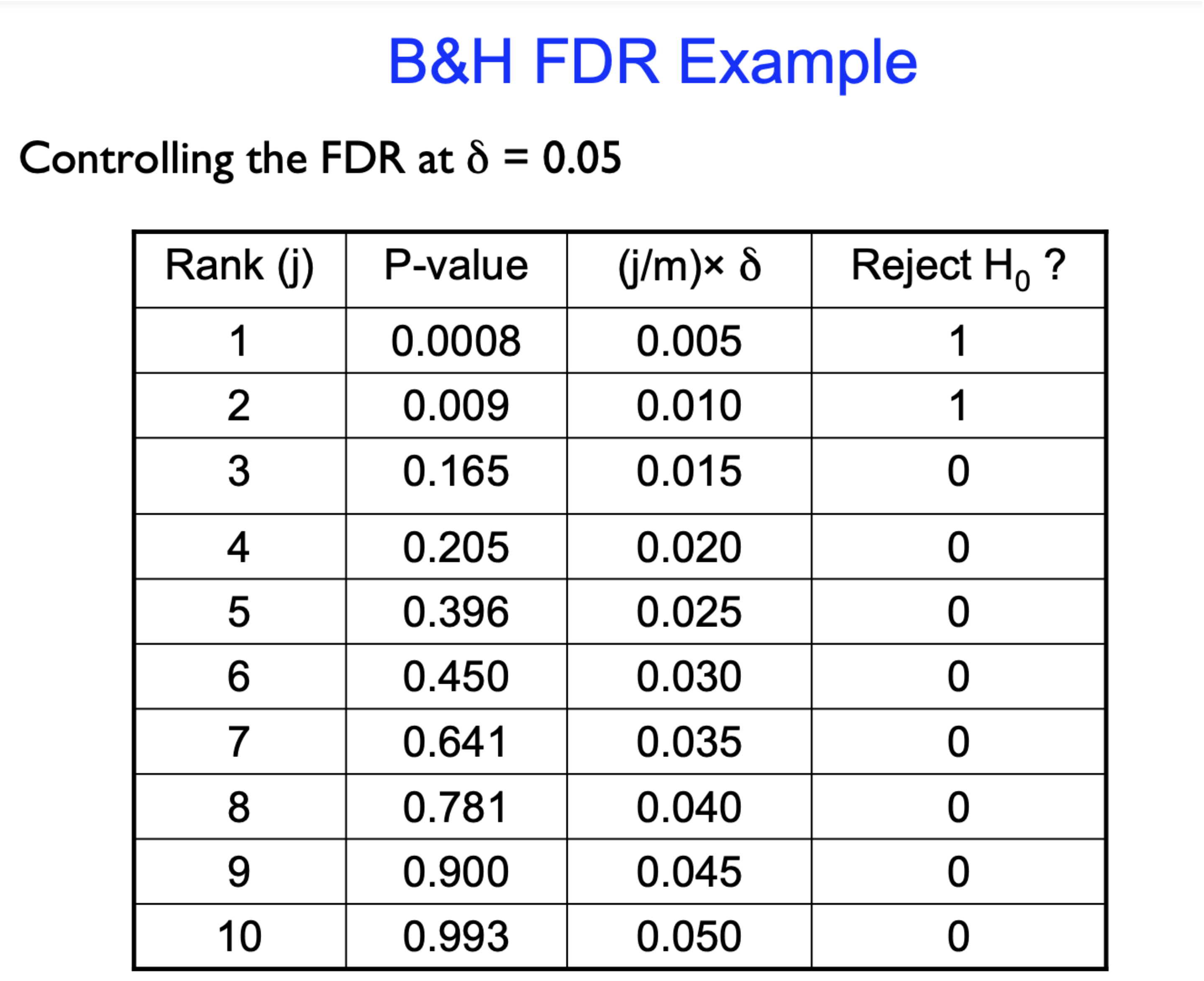
Benjamini-Hochberg critical value = $(i\ /\ \mathrm{m})\cdot\mathrm{Q}$

- Step 1: Rank the p values from smallest to largest

- Step 2: Compare against the B-H critical value.
- i = rank, m = number of tests, Q = chosen FDR

- Step 3: The largest p value that is < the critical value is significant and so are all the p values in your list that are less than this p value

Controlling for False Discovery Rate (FDR) - Benjamini-Hochberg procedure



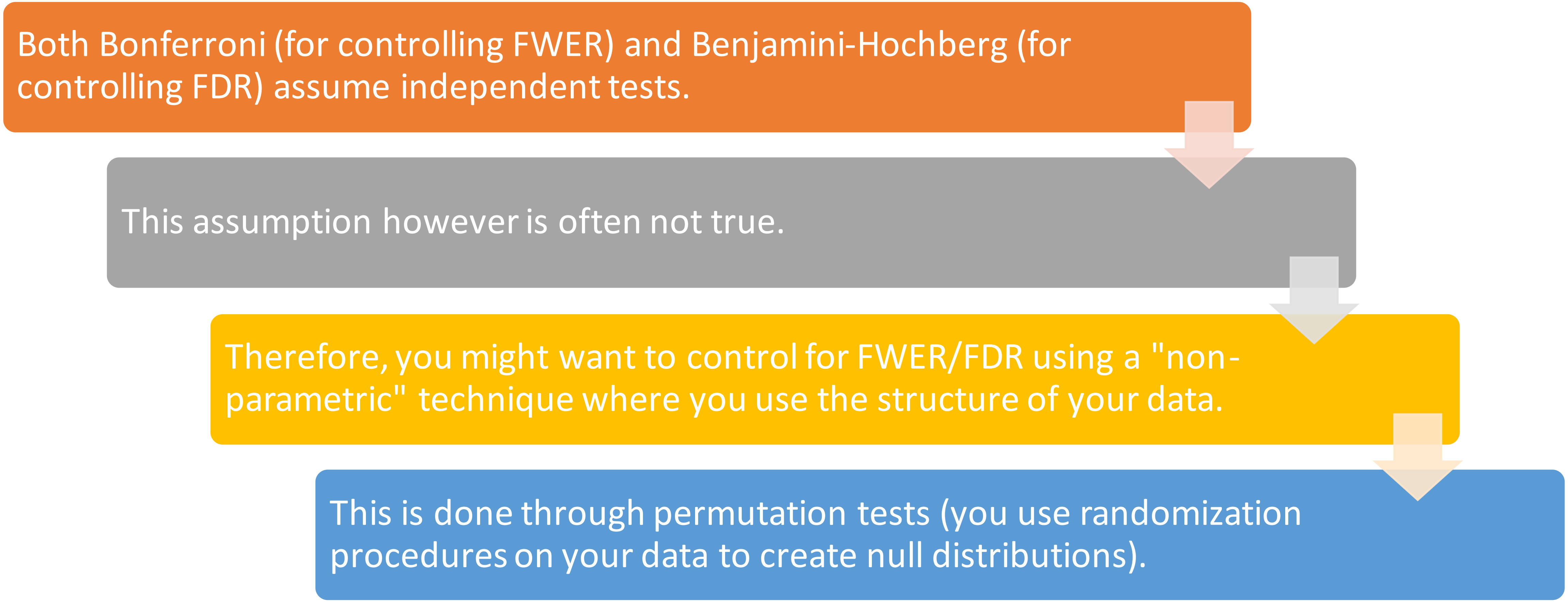
B&H FDR Example

Controlling the FDR at δ = 0.05

Rank (j)	P-value	(j/m)× δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture10.pdf

The independence assumption



Both Bonferroni (for controlling FWER) and Benjamini-Hochberg (for controlling FDR) assume independent tests.

This assumption however is often not true.

Therefore, you might want to control for FWER/FDR using a "non-parametric" technique where you use the structure of your data.

This is done through permutation tests (you use randomization procedures on your data to create null distributions).

The general idea behind using permutation-based simulations for controlling FWER

- m tests --> m "uncorrected" p-values
- Randomize the experiment 1000+ times
- Conduct all m tests each time – we know that all m null hypotheses are true in the randomizations (by construction)
- Determine from the simulations, the target p-value threshold for which you get 5% significant results across the 1000+ simulations -- i.e., calculate the proportion of simulations in which you get at least one significant test.
- Use this threshold now as the criterion for the uncorrected p-values in the true data, all those below the criterion are significant.

A final note before we look at permutation-based procedures in detail

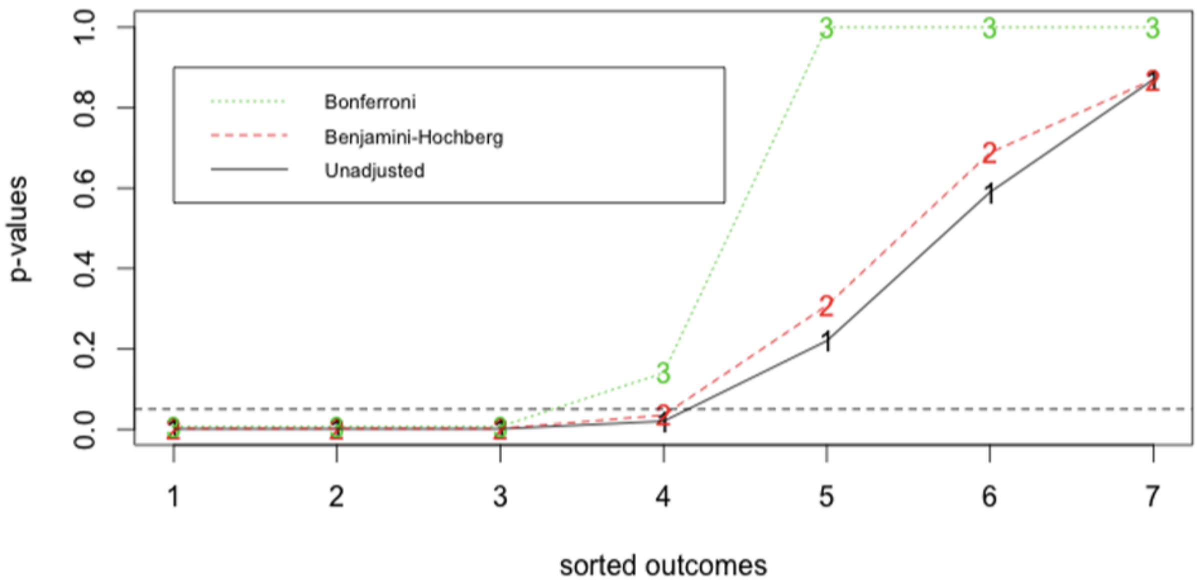
The answer will be similar to parametric methods if the m tests are independent.

The simulation-based threshold **will be less stringent** than the parametric versions **as more of the m tests become less independent from each other.**

You should try this out with your data and get an intuitive grasp.

1. The following array *p* consists of the observed significance values for multiple correlation tests.
   p = [0.0050 0.0010 0.0100 0.0005 0.0009 0.0400 0.0560 0.0500 0.0480 0.0130 0.0370 0.0430 0.0020 0.0250 0.1100 0.0700 0.0800]
   Apply both Bonferroni and Benjamini-Hochberg correction and create a graph as shown below with the observed/unadjusted p-values, Bonferroni corrected and BH-corrected ones for all the tests sorted and comment on the results (ensure you also plot the black dashed line that represents an alpha level of .05). The relevant function is **p.adjust** in R. The Bonferroni method is known to be a more conservative approach. Do the results of the correction support that?

   .


Bonferroni
Benjamini-Hochberg
Unadjusted
p-values
sorted outcomes