Genomics Data :

Test multiplicity

Introduction to Hypothesis Testing

Why do hypothesis testing?

Hypothesis testing = making decisions with a **finite sample of noisy observations**

E.g. Do patients respond better to treatment A than treatment B?

- **Finite sample** : cost, ethics, time…
- **Noisy observations** : patient responses depend on factors outside of treatment

Motivating Example : Coin Tossing

I want to test whether a coin is *fair* i.e. 50% probability of both heads and tails

I flip the coin 100 times and observe a sequence of heads and tails :

HTTTHHHTTHTHTT…..

And I count the number of heads and tails

Heads	**Tails**
60	40

Do we have enough information to make a conclusion?

The null distribution

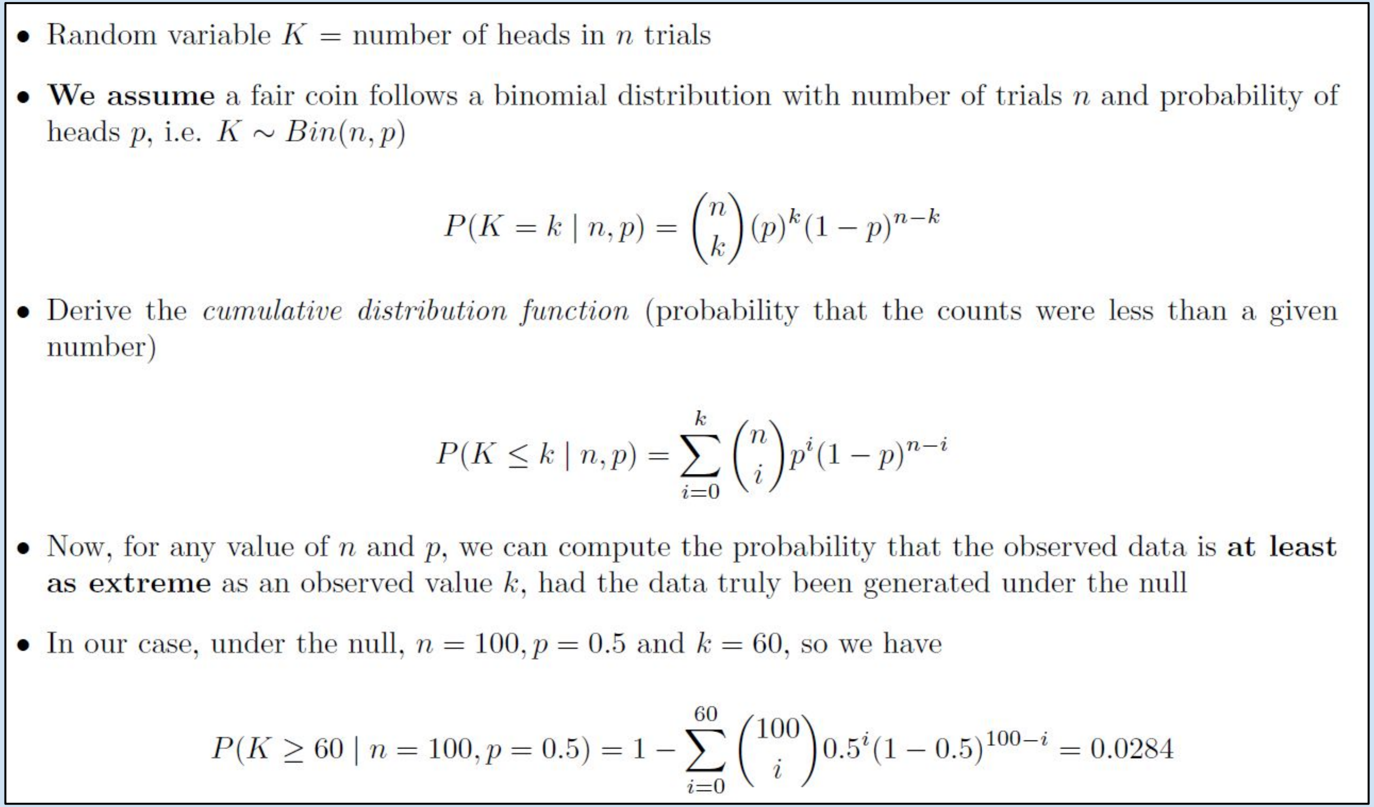
- Even if the coin is fair, we expect random sampling differences
- To conclude if the coin is fair or not based on the observed data, we need to specify what we would ***expect*** **to happen if the coin was fair**
- This is called the **null distribution**, and tells us what the data should look like under the **null hypothesis**

Null hypothesis : prob(heads) = 0.5
Alternative hypothesis : prob(heads) != 0.5

How to compute a null distribution?

- **Analytically :** assume a **theoretical parametric distribution** of the data under the null
  - I.e. the “pen and paper” method
  - Practical, easy to recompute for a range of parameters
  - Requires modelling assumptions, not always simple to compute
- **Monte-Carlo Simulation :** simulate data under the null hypothesis
  - Intuitive, does not require modelling assumptions
  - Long to compute

Computing the null distribution *analytically*



- Random variable K = number of heads in n trials
- **We assume** a fair coin follows a binomial distribution with number of trials n and probability of heads p, i.e. $K \sim Bin(n,p)$

$$P(K = k \mid n, p) = \binom{n}{k} (p)^k (1-p)^{n-k}$$

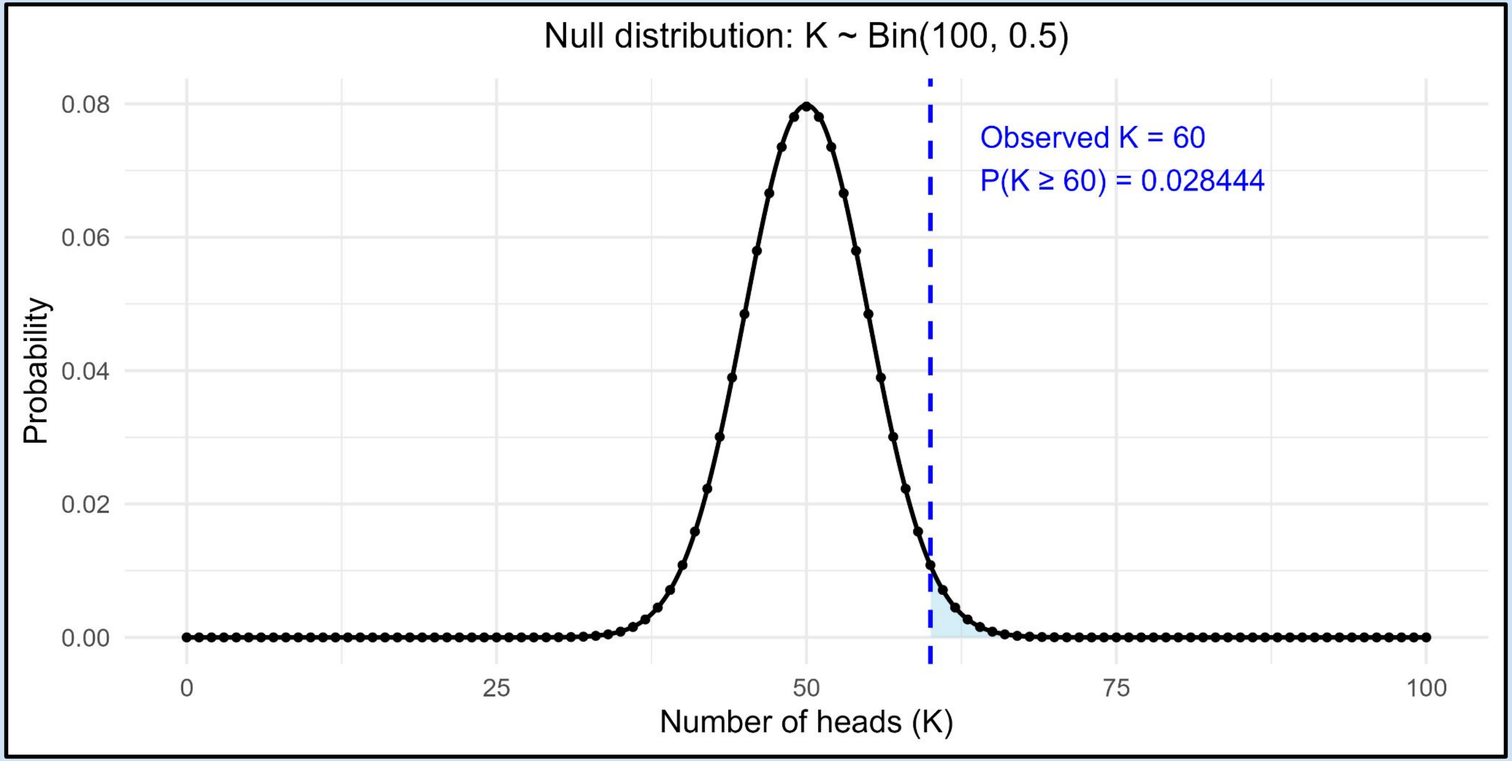
- Derive the *cumulative distribution function* (probability that the counts were less than a given number)

$$P(K \leq k \mid n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

- Now, for any value of n and p, we can compute the probability that the observed data is **at least as extreme** as an observed value k, had the data truly been generated under the null
- In our case, under the null, $n = 100, p = 0.5$ and $k = 60$, so we have

$$P(K \geq 60 \mid n = 100, p = 0.5) = 1 - \sum_{i=0}^{60} \binom{100}{i} 0.5^i (1-0.5)^{100-i} = 0.0284$$

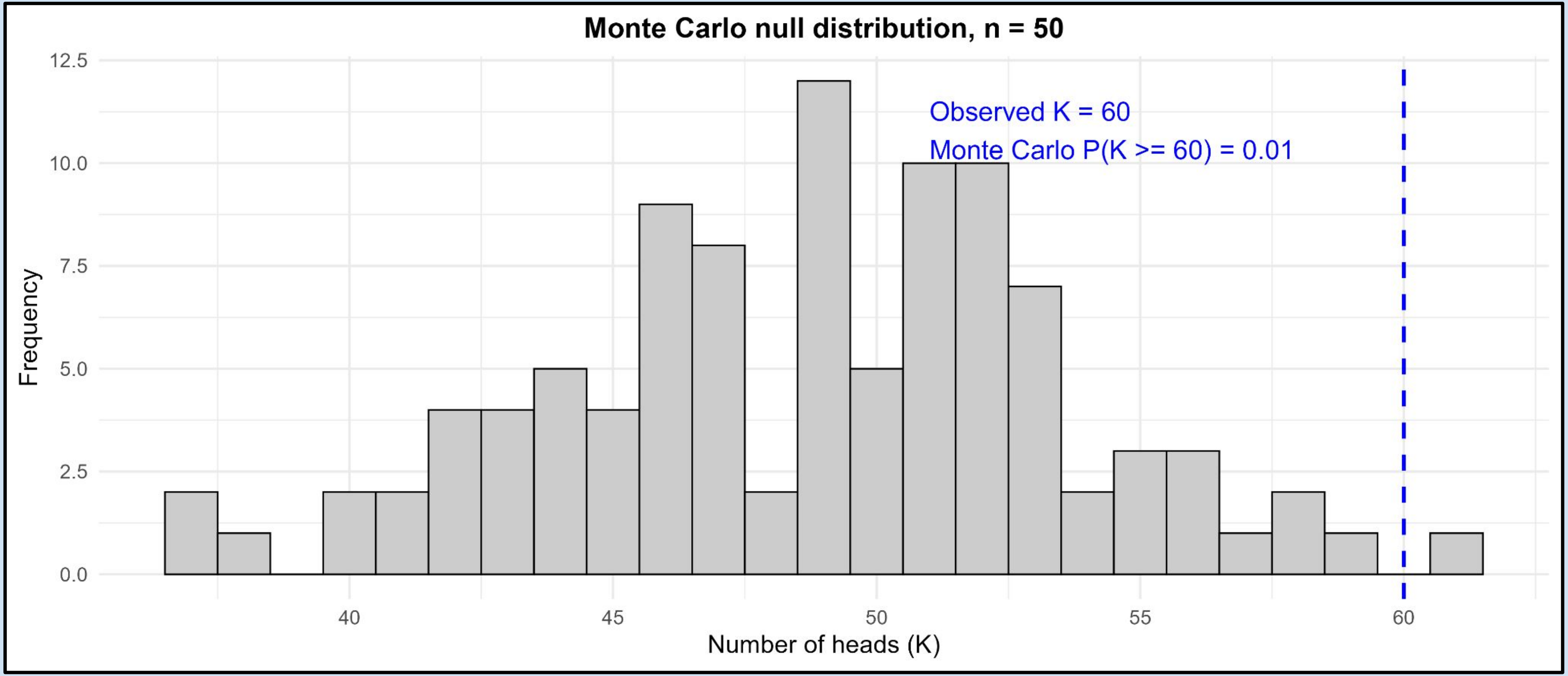
Computing the null distribution *analytically*



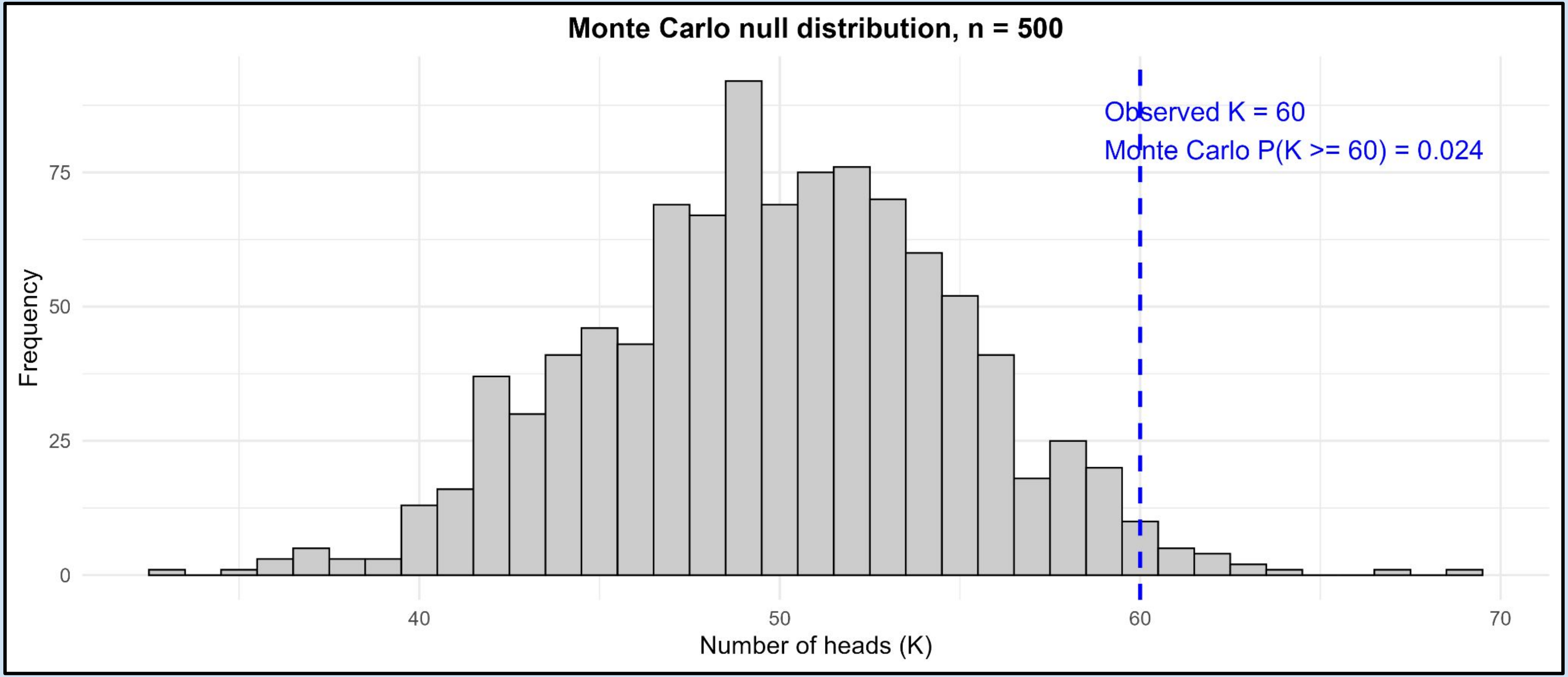
Computing the null with *Monte-Carlo Simulation*

- **Idea** : repeatedly simulate realisations of data from the null
  - I.e. in R, rbinom(1, size = 100, prob = 0.5)
- Estimate the extremeness of the observed data by computing the **proportion of simulated values greater than the observed one**
- The more realisations we do, the closer we get to the analytical solution

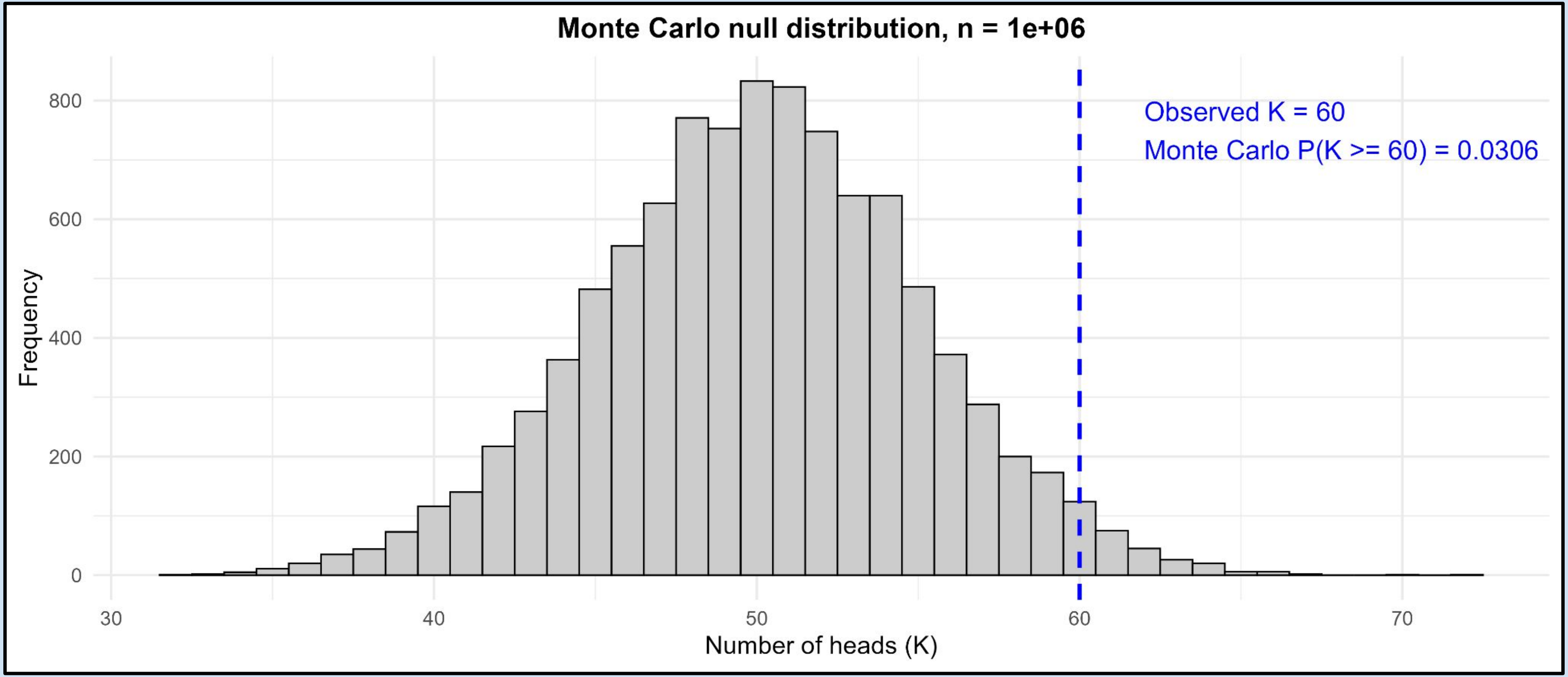
Computing the null with *Monte-Carlo Simulation*



Computing the null with *Monte-Carlo Simulation*

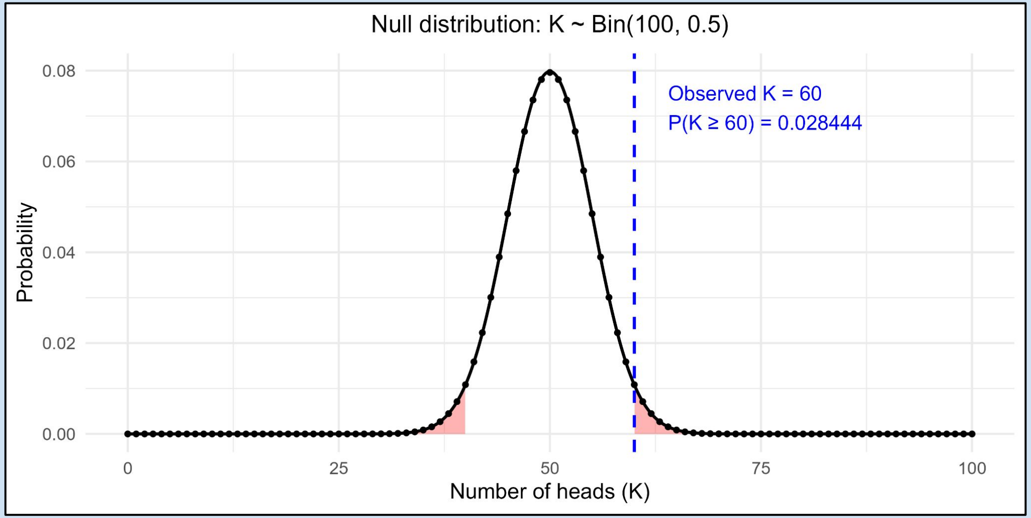


Computing the null with *Monte-Carlo Simulation*

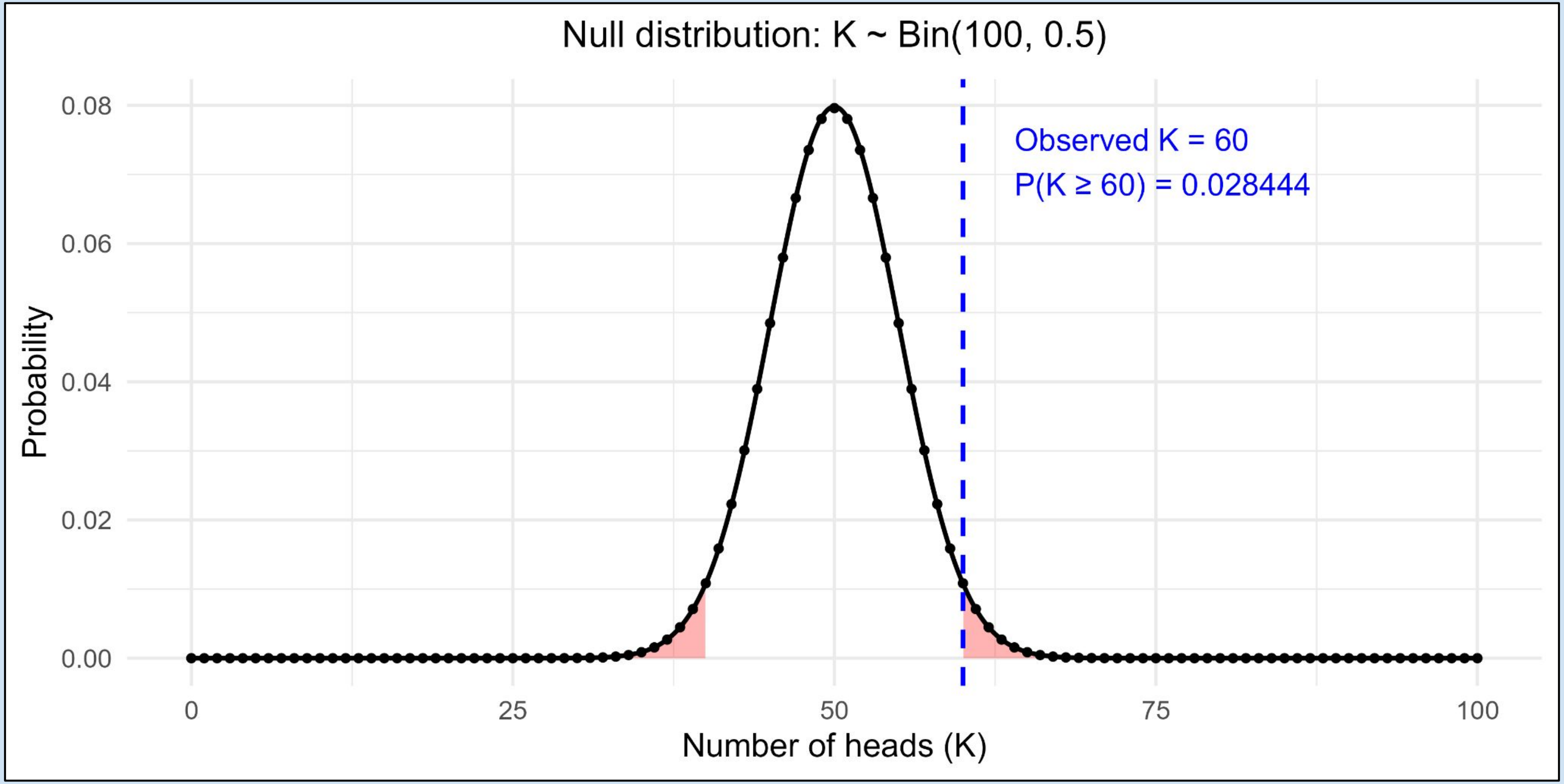


Determining a rejection region

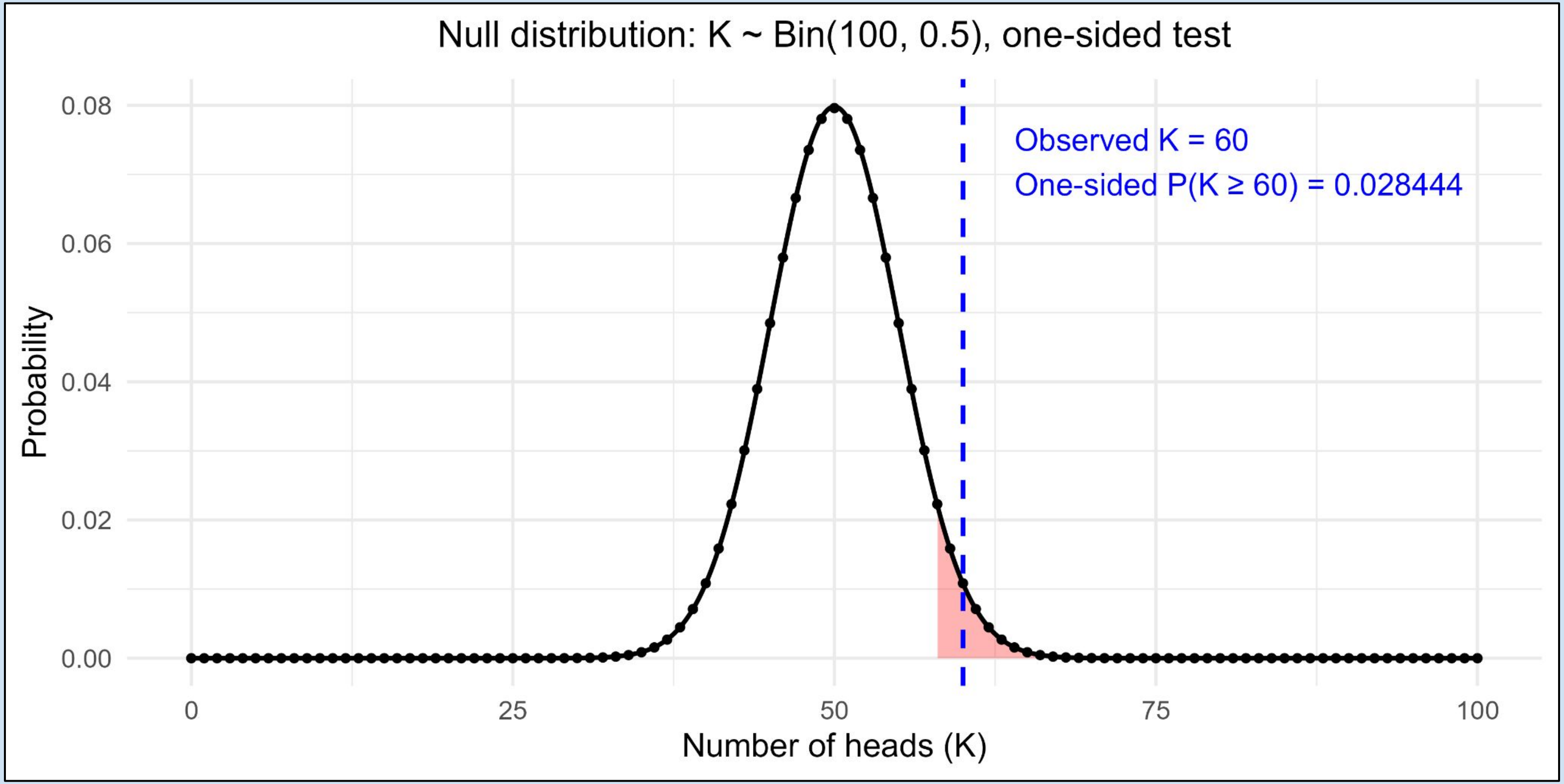
- If our observed value falls in the rejection region, we reject the null
- The rejection region is determined by the significance level of the test
  - I.e. how extreme should the observed data need to be for me to reject the null?



Two-sided test, significance level = 5%



One-sided test, significance level = 5%



5 steps to hypothesis testing

1. Test statistic
   - **A summary of the data** used to make the decision
   - E.g. proportion of heads
2. Null hypothesis and distribution
   - Analytically or with simulation
3. Rejection region
   - Region of the null distribution for which we consider the test significant
4. Observe data
5. Decision
   - Either reject the null, or do not reject the null
   - We cannot **prove** a null with hypothesis testing!

Types of error

	Null true	**Null false**
Reject null	False positive (T1 error)	True Positive
Do not reject null	True negative	False negative (T2 error)

There is a tradeoff between false positives and false negatives

- For example, I could make a medical test which gives negative for every patient.
- I would never identify any false positives! Type 1 error rate = 0
- However, I would never identify any true positives either, i.e. power = 0

Pitfalls with p-values : hacking, HARKing

P-hacking

- **Torturing the data until a significant p-value is found**
- E.g. in the coin example, we could consider different test statistics like number of consecutive heads, or we could take different subsets of the data
- Classical example in regression modelling : specifying multiple models

HARKing

- Hypothesising after the results are known
- Changing the null hypothesis after investigating the data

Multiple Testing Problem

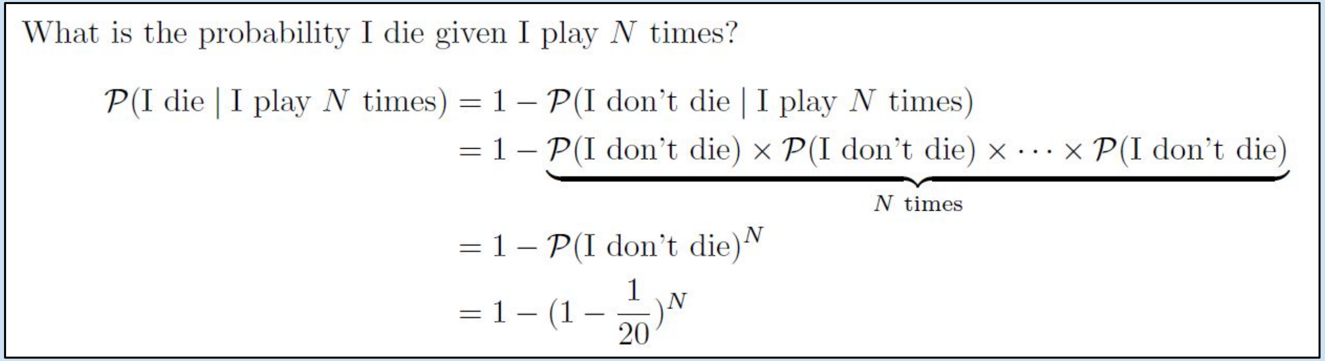
Motivating Example : Russian Roulette

Imagine a gun with **20 chambers**, where **one is loaded** with a bullet

I am going to randomly spin the chamber and pull the trigger

I would like to know **the probability that I die if I repeat this process** a certain number of times

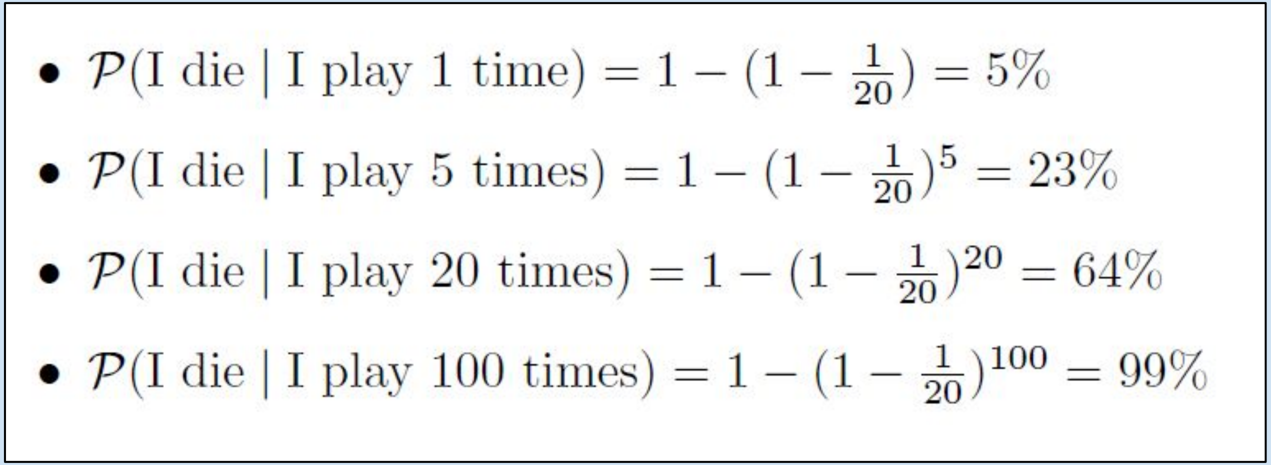
Motivating Example : Russian Roulette



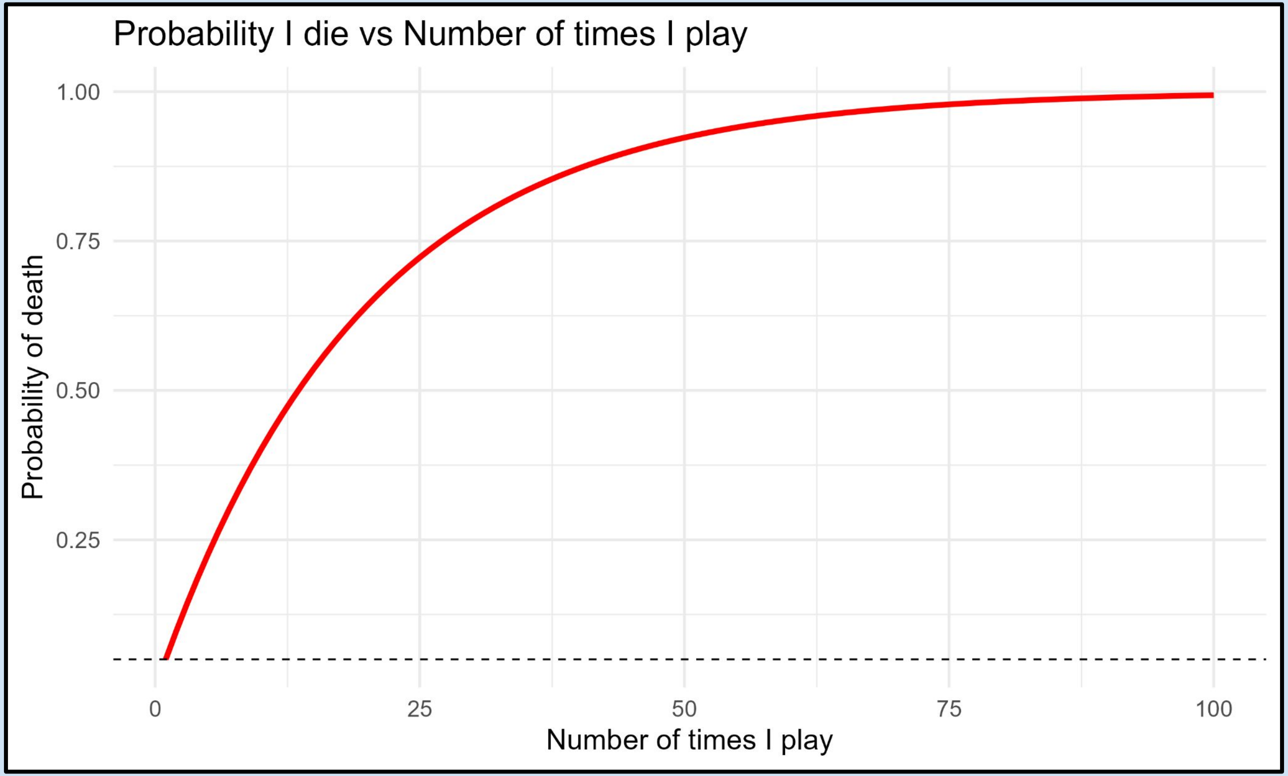
What is the probability I die given I play N times?

$$\begin{aligned}
\mathcal{P}(\text{I die} \mid \text{I play } N \text{ times}) &= 1 - \mathcal{P}(\text{I don't die} \mid \text{I play } N \text{ times}) \\
&= 1 - \underbrace{\mathcal{P}(\text{I don't die}) \times \mathcal{P}(\text{I don't die}) \times \cdots \times \mathcal{P}(\text{I don't die})}_{N \text{ times}} \\
&= 1 - \mathcal{P}(\text{I don't die})^N \\
&= 1 - (1 - \frac{1}{20})^N
\end{aligned}$$

Motivating Example : Russian Roulette

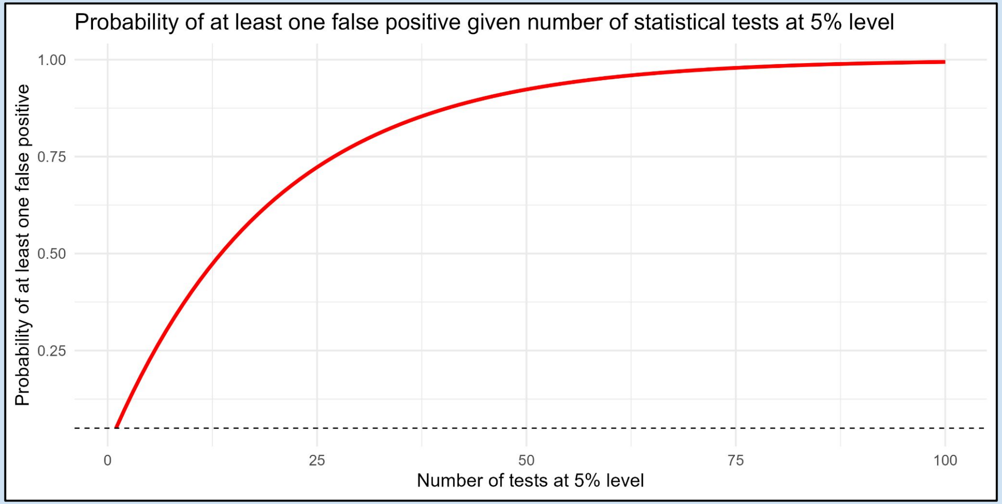


- $\mathcal{P}(\text{I die} \mid \text{I play 1 time}) = 1 - (1 - \frac{1}{20}) = 5\%$
- $\mathcal{P}(\text{I die} \mid \text{I play 5 times}) = 1 - (1 - \frac{1}{20})^5 = 23\%$
- $\mathcal{P}(\text{I die} \mid \text{I play 20 times}) = 1 - (1 - \frac{1}{20})^{20} = 64\%$
- $\mathcal{P}(\text{I die} \mid \text{I play 100 times}) = 1 - (1 - \frac{1}{20})^{100} = 99\%$


Probability I die vs Number of times I play
1.00
0.75
0.50
0.25
Probability of death
0
25
50
75
100
Number of times I play

Hypothesis testing is scientific roulette!

- In traditional hypothesis testing, significance level is typically 5%
    - This means, given all my testing assumptions are met, I expect 5% of results under the null hypothesis to be called positive (i.e. false positives)
- more tests, more likely to have at least one false positive



Types of error : notation

Null hypothesis Test	True	False	Total
Non-rejected	U	T	W
Rejected	V	S	R
Total	m0	m-m0	m

Family-Wise Error Rate

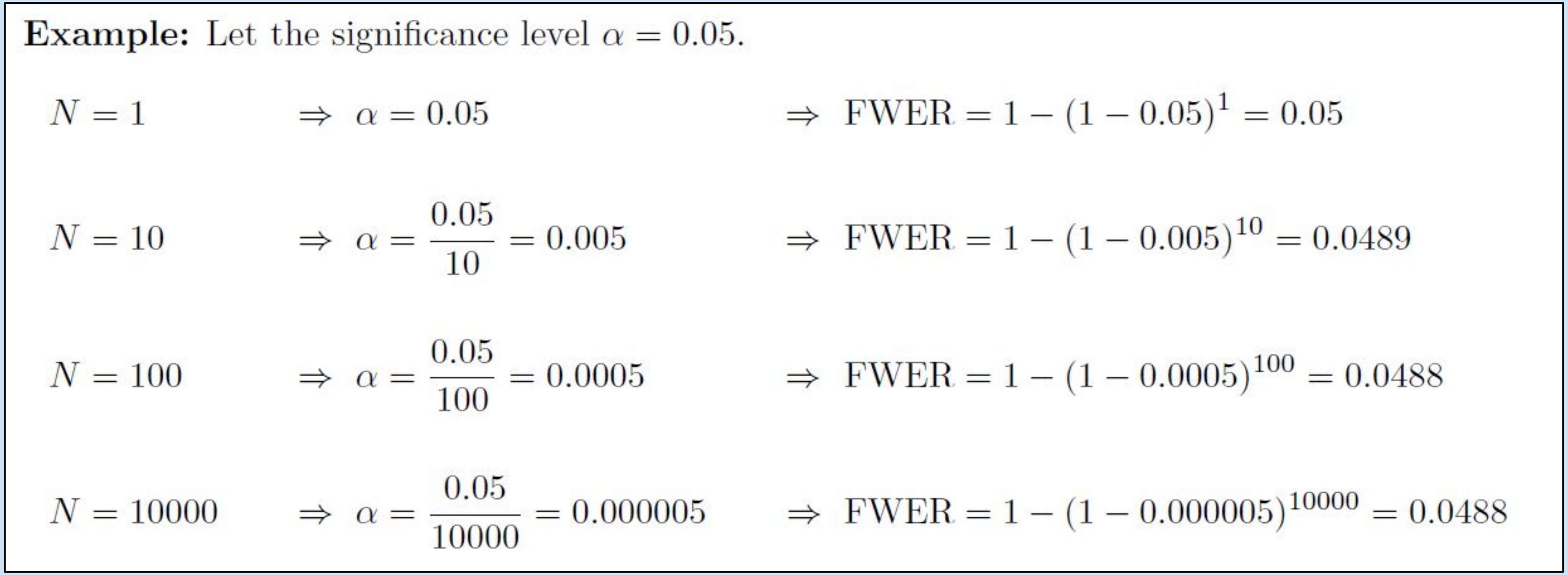
- What we just computed is called the ***Family-Wise Error Rate :*** the probability that at least one of my positive test results is false - P(V > 0)
- How could we ***control*** this quantity i.e. ***bound*** it by an upper limit?

Null hypothesis **Test**	**True**	**False**	**Total**
Non-rejected	U	T	W
Rejected	V	S	R
Total	m0	m-m0	m

Bonferroni Correction

Idea : divide the significance level by the number of tests N

- I.e. for m tests, reject hypothesis if p value < α/N



Example: Let the significance level $\alpha = 0.05$.

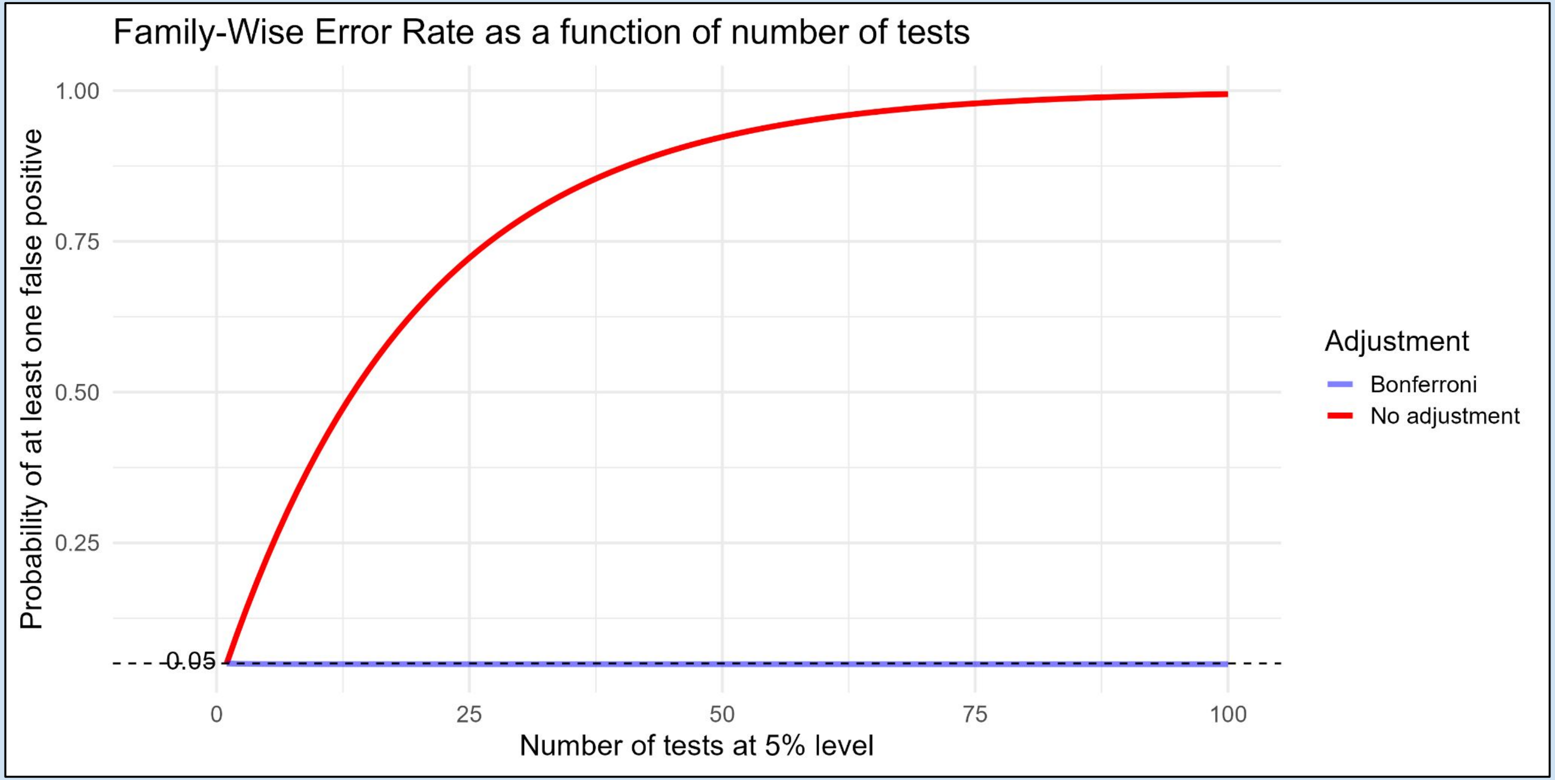
$$N = 1 \quad \Rightarrow \alpha = 0.05 \quad \Rightarrow \text{FWER} = 1 - (1 - 0.05)^1 = 0.05$$

$$N = 10 \quad \Rightarrow \alpha = \frac{0.05}{10} = 0.005 \quad \Rightarrow \text{FWER} = 1 - (1 - 0.005)^{10} = 0.0489$$

$$N = 100 \quad \Rightarrow \alpha = \frac{0.05}{100} = 0.0005 \quad \Rightarrow \text{FWER} = 1 - (1 - 0.0005)^{100} = 0.0488$$

$$N = 10000 \quad \Rightarrow \alpha = \frac{0.05}{10000} = 0.000005 \quad \Rightarrow \text{FWER} = 1 - (1 - 0.000005)^{10000} = 0.0488$$

Bonferroni Correction

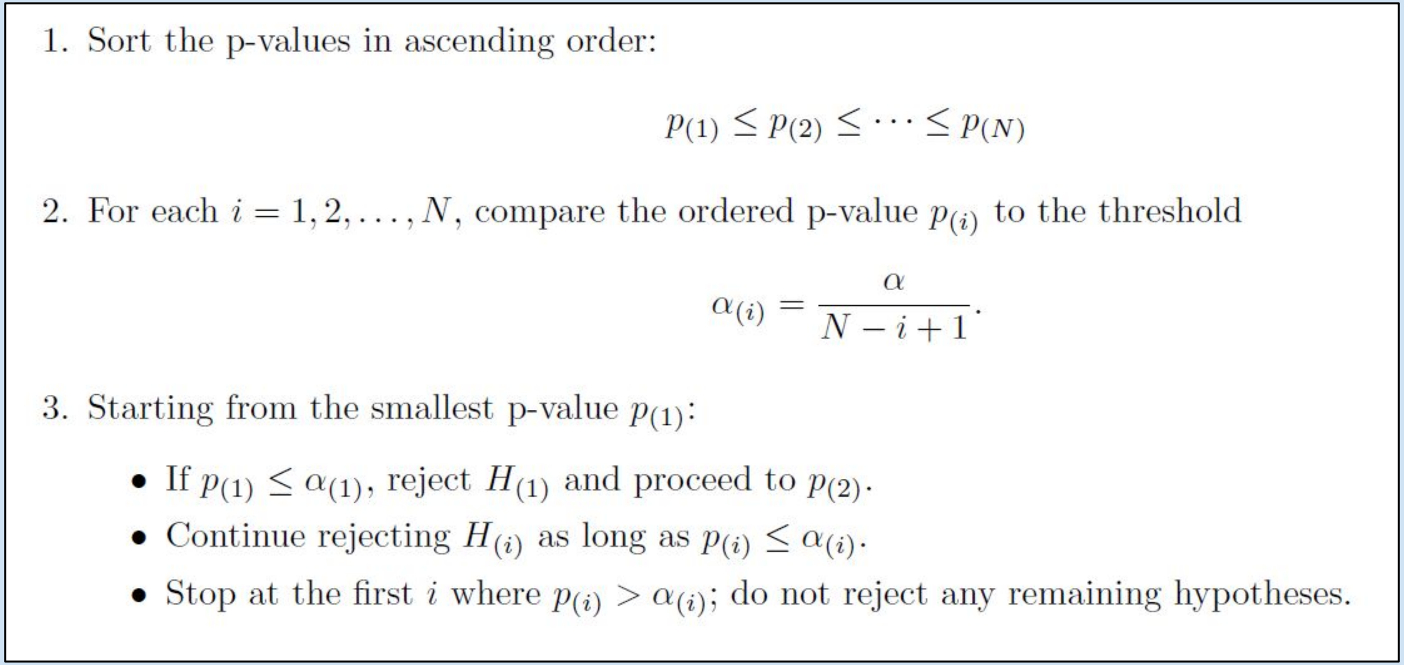


Bonferroni Correction

- Strictly controls the FWER
- However, very conservative when N is large
- For example, when N = 10,000, p-values must be smaller than 0.000005 to reject
- Also, every test has the same significance threshold, regardless of level of evidence
- Idea : control the FWER but use ***sequential threshold*** which becomes less strict as the p-values get larger

Holm Correction

Idea: relax the significance threshold as the p-values get larger



1. Sort the p-values in ascending order:

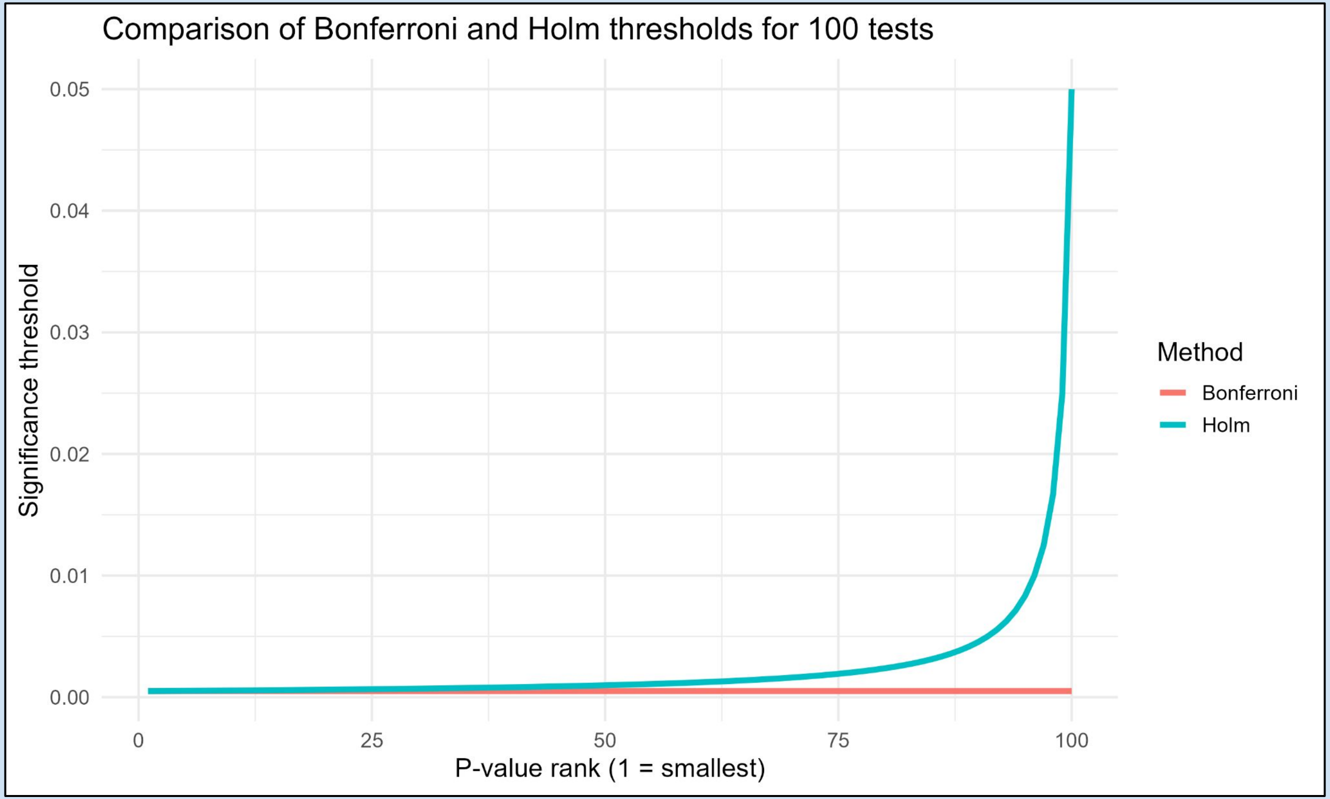
$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$$

2. For each $i = 1, 2, \ldots, N$, compare the ordered p-value $p_{(i)}$ to the threshold

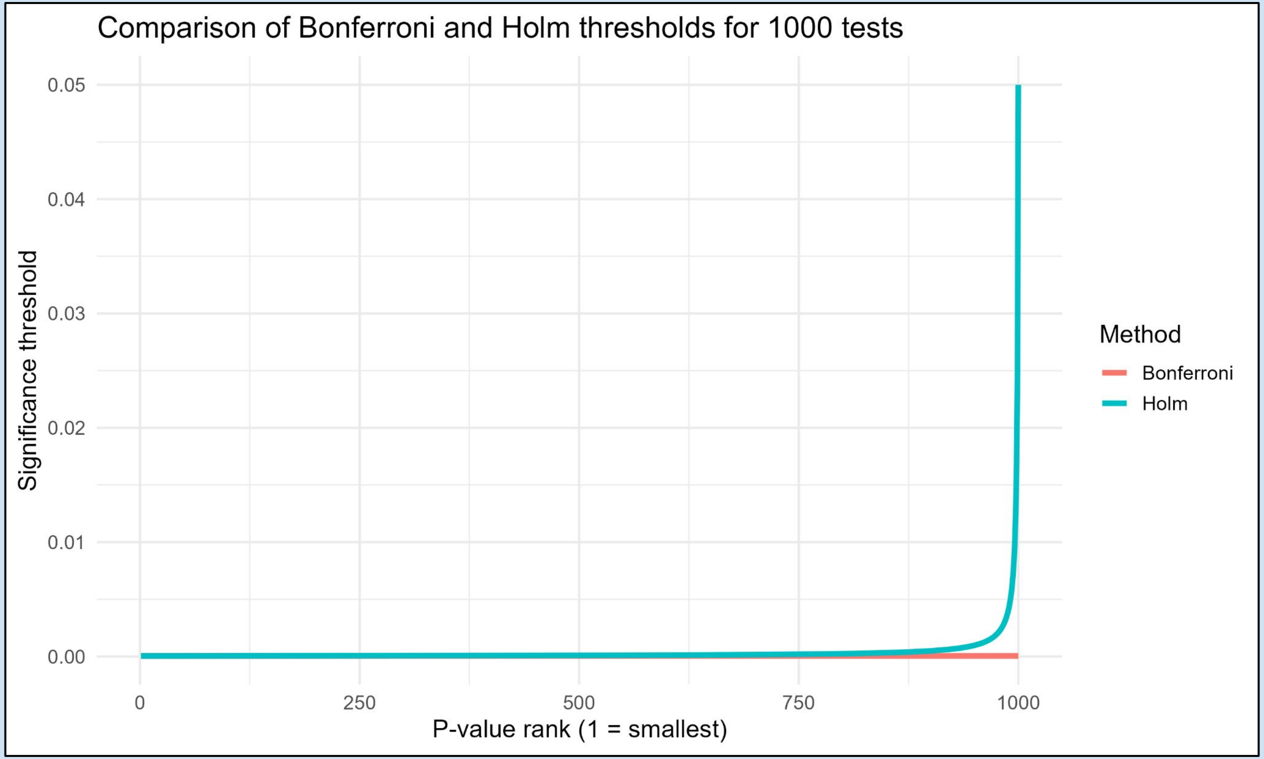
$$\alpha_{(i)} = \frac{\alpha}{N - i + 1}.$$

3. Starting from the smallest p-value $p_{(1)}$:
   - If $p_{(1)} \leq \alpha_{(1)}$, reject $H_{(1)}$ and proceed to $p_{(2)}$.
   - Continue rejecting $H_{(i)}$ as long as $p_{(i)} \leq \alpha_{(i)}$.
   - Stop at the first i where $p_{(i)} > \alpha_{(i)}$; do not reject any remaining hypotheses.

Holm vs Bonferroni correction - 100 tests



Holm vs Bonferroni correction - 1,000 tests



Summary on controlling the FWER

- FWER = Probability of at least one false positive result
- This quickly goes to 1 as we increase the number of tests
- Bonferroni *controls* the FWER by dividing the significance level by the number of tests
- Holm is less conservative as it uses thresholds which increase with the rank of the p-value

Summary on controlling the FWER

- FWER = Probability of at least one false positive result
- This quickly goes to 1 as we increase the number of tests
- Bonferroni *controls* the FWER by dividing the significance level by the number of tests
- Holm is less conservative as it uses thresholds which increase with the rank of the p-value
- FWER methods are conservative as they control the probability to obtain at least one FP
- Perhaps we don't care about a few FPs as long as they do not dominate our significant results?

False Discovery Rate

False Discovery Rate

FDR = expected proportion of significant results which are false

- E(V/R)
- V/R is called the **false discovery proportion (FDP)**
- Conceptually this is similar to **positive predictive value** in epidemiology
- I.e. if I have a positive test, what is the probability that it is true?

	True	**False**	**Total**
Non-rejected	U	T	W
Rejected	V	S	R
Total	m0	m-m0	m

An example of FDR

- Imagine a test which has **100% power** at significance level **5%**
- It **identifies every true positive**
- 5% of true negatives will be falsely called positive

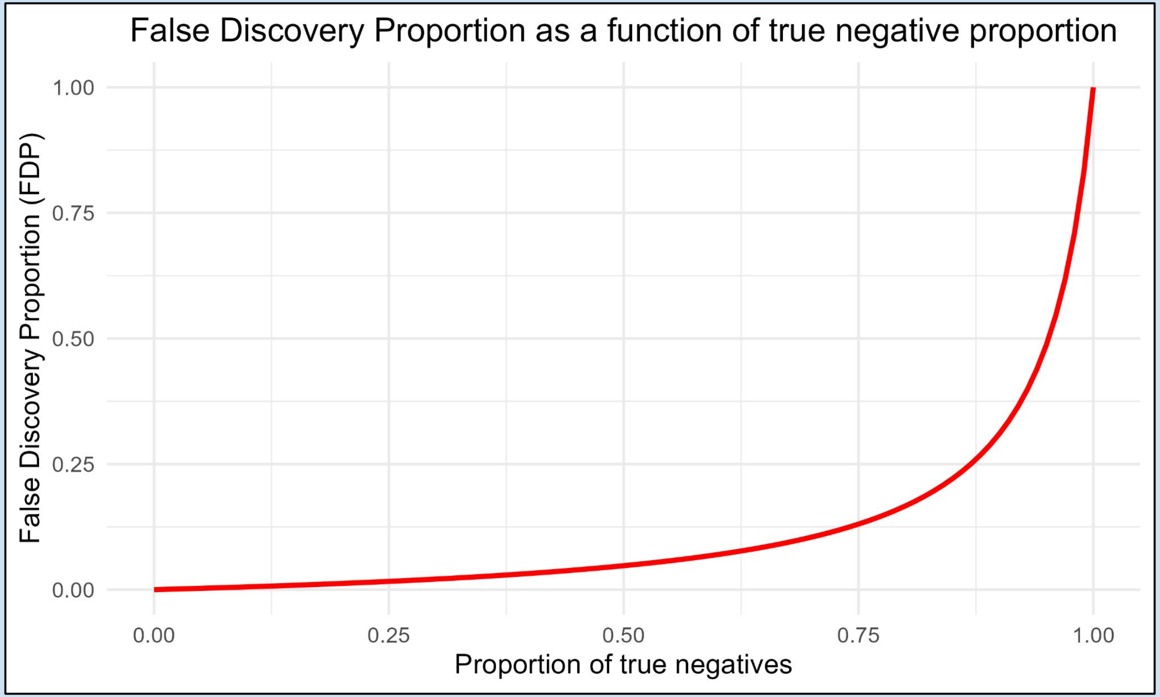
Let’s say I test 50 variables, 10 of which are true positives

- I identify 10 true positives and 40*0.05 = 2 false positives
- FDP = 2/(10+2) = 16.7%

An example of FDR

Now I test 1000 variables, still with 10 true positives

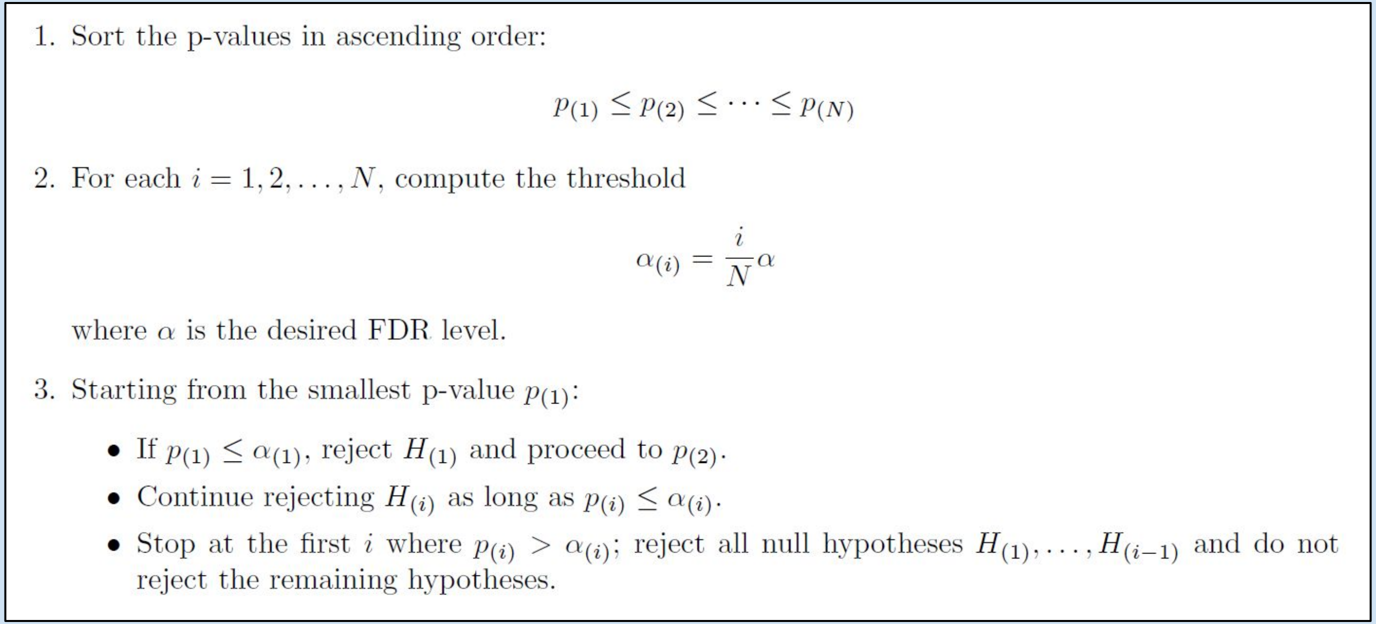
- I identify 10 true positives and 990 * 0.05 = 50 false positives
- Now FDP = 50/(50+10) = 83%
- My results are becoming a bit worthless…



When we have a lot of true negatives relative to true positives, false positives will dominate our results!

Benjamini-Hochberg

- BH is a method to control the FDR
- Similar to Holm, it uses sequential thresholds on the ranked p-values



1. Sort the p-values in ascending order:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$$

2. For each $i = 1, 2, \ldots, N$, compute the threshold

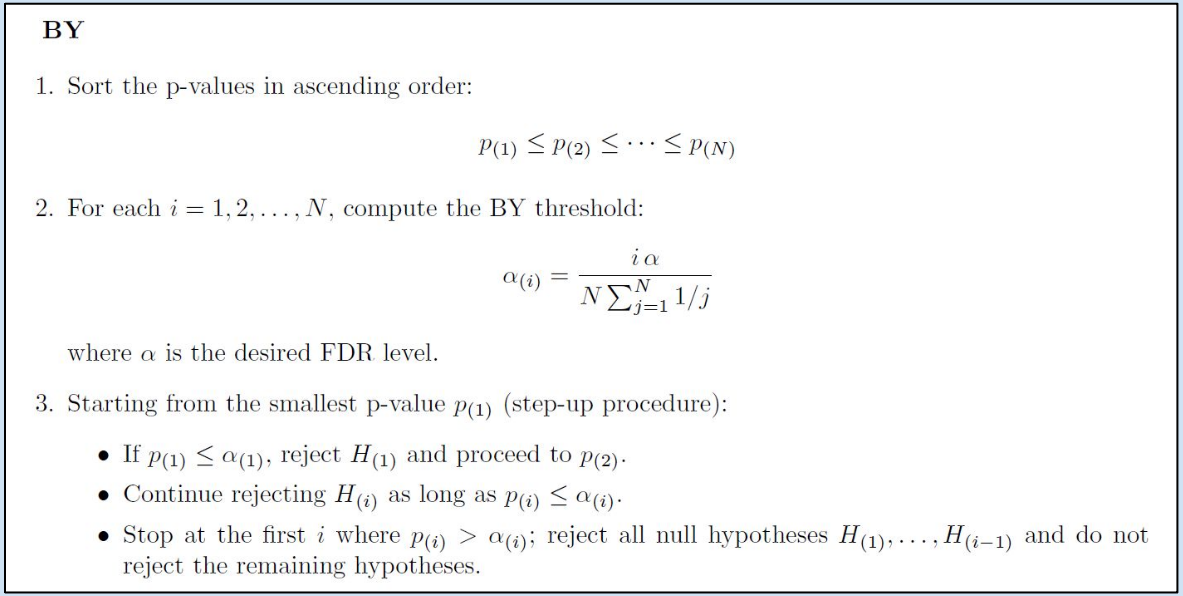
$$\alpha_{(i)} = \frac{i}{N}\alpha$$

   where α is the desired FDR level.

3. Starting from the smallest p-value $p_{(1)}$:
   - If $p_{(1)} \leq \alpha_{(1)}$, reject $H_{(1)}$ and proceed to $p_{(2)}$.
   - Continue rejecting $H_{(i)}$ as long as $p_{(i)} \leq \alpha_{(i)}$.
   - Stop at the first i where $p_{(i)} > \alpha_{(i)}$; reject all null hypotheses $H_{(1)}, \ldots, H_{(i-1)}$ and do not reject the remaining hypotheses.

Benjamini-Yekutieli

- Problem : B-H assumes independent tests
- B-Y fixes this by assuming some dependencies among tests
- This procedure is more conservative than B-H



BY

1. Sort the p-values in ascending order:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$$

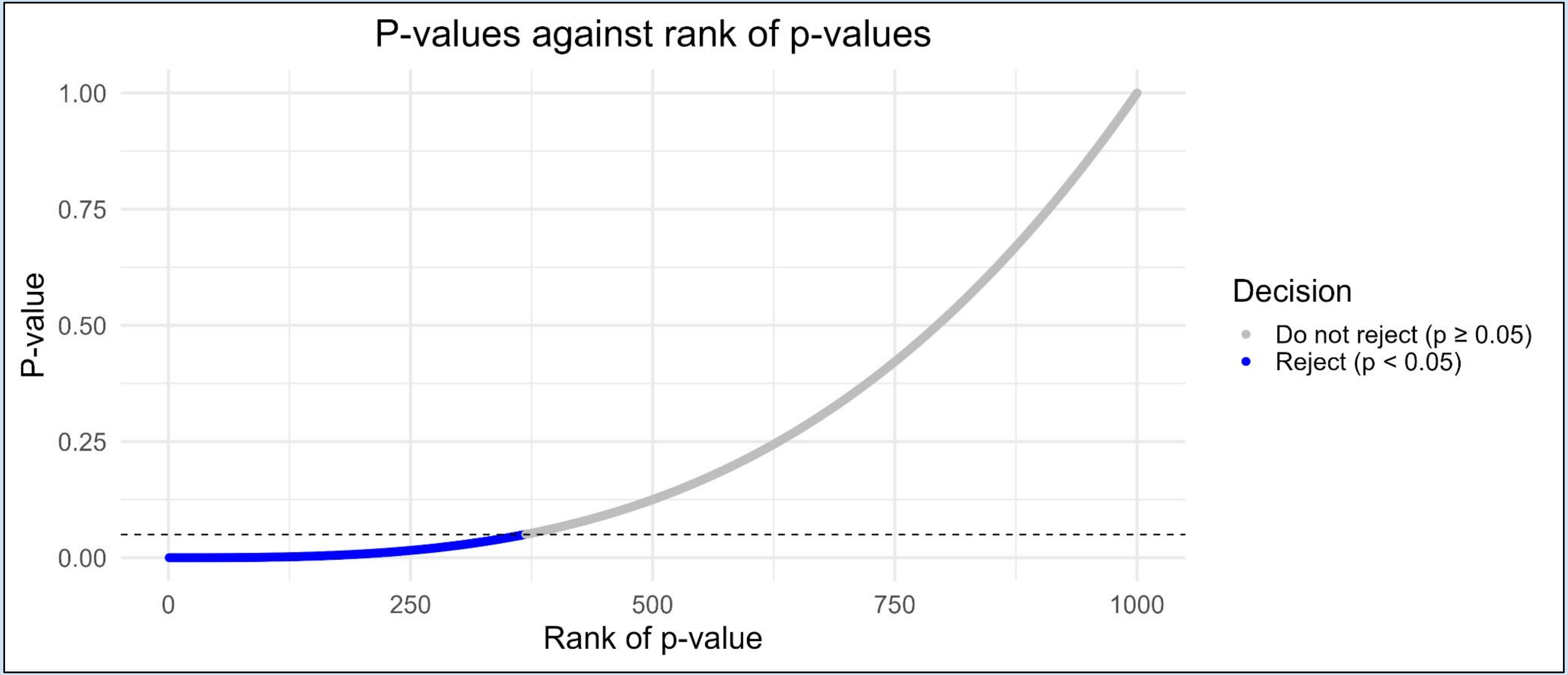
2. For each $i = 1, 2, \ldots, N$, compute the BY threshold:

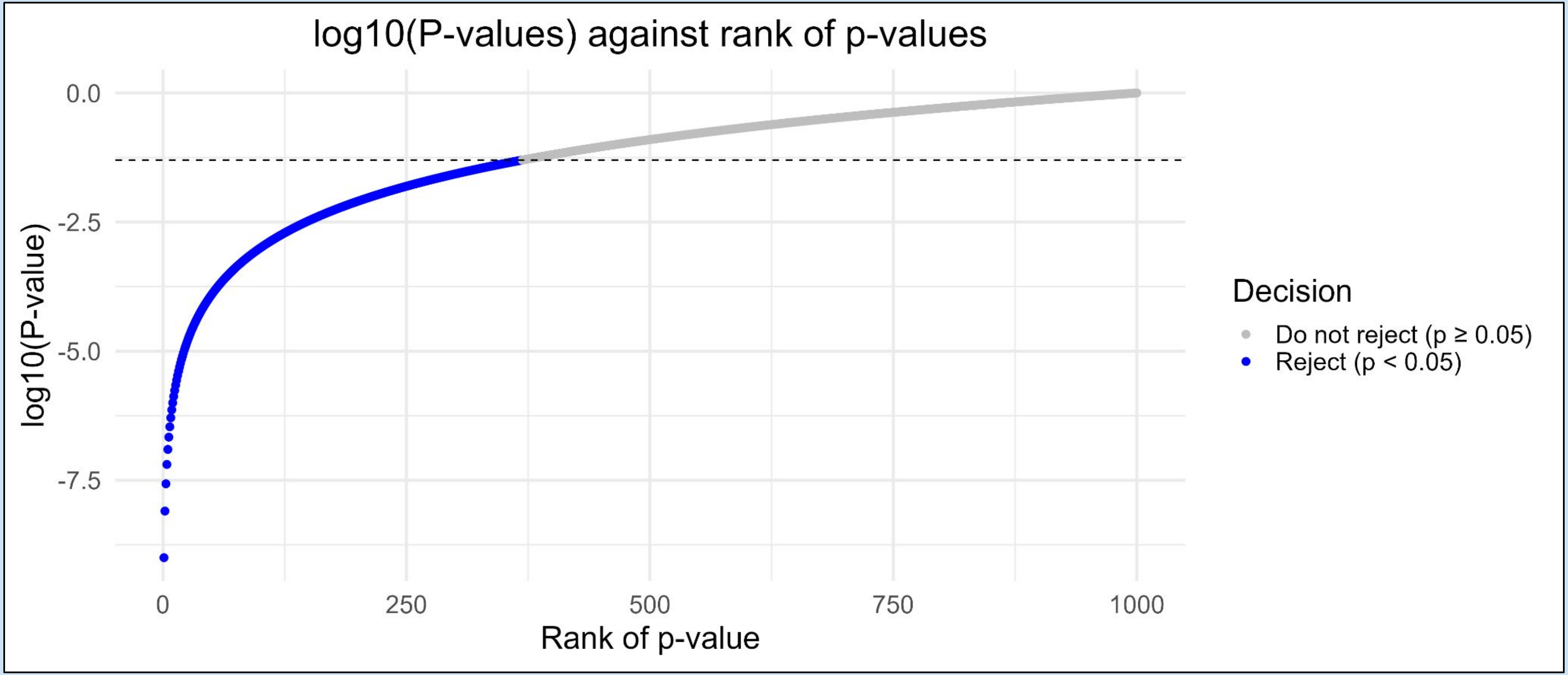
$$\alpha_{(i)} = \frac{i\,\alpha}{N \sum_{j=1}^{N} 1/j}$$

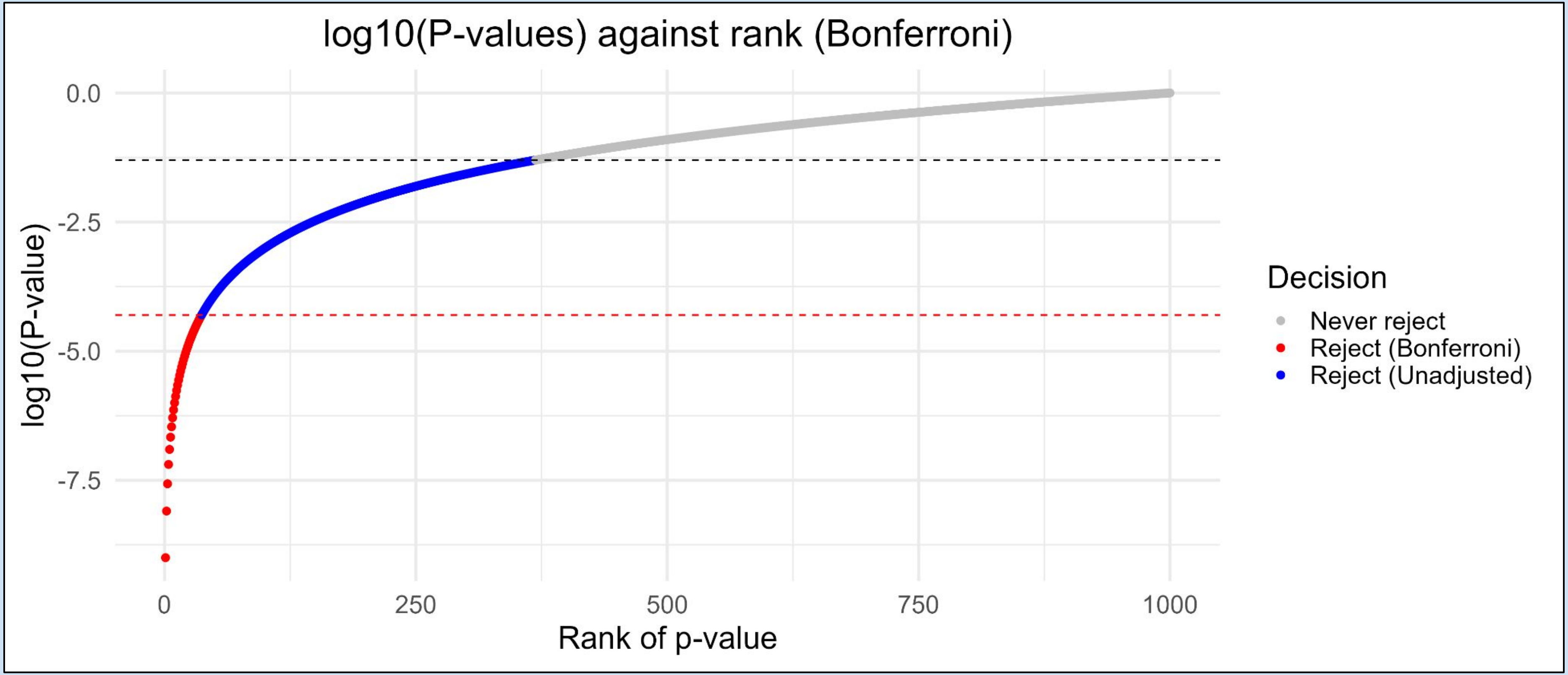
where α is the desired FDR level.

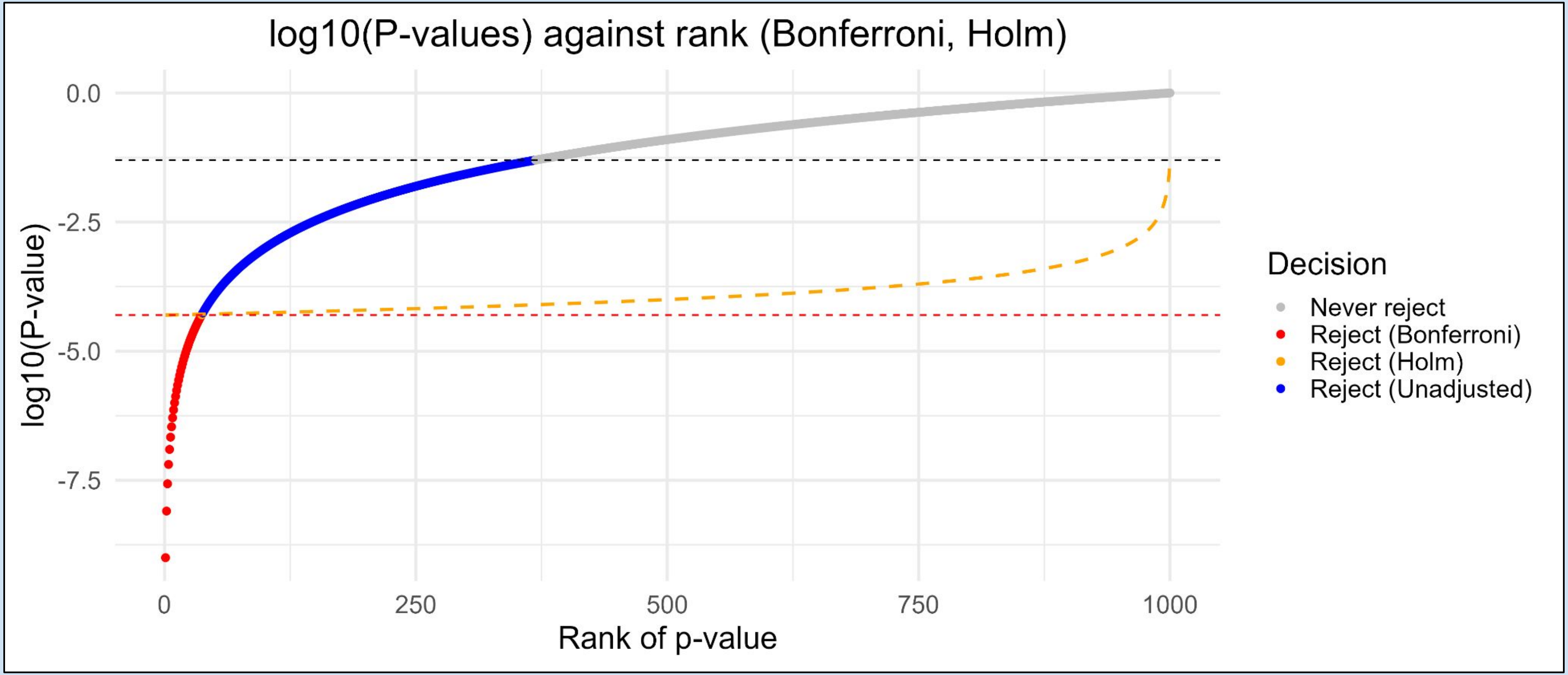
3. Starting from the smallest p-value $p_{(1)}$ (step-up procedure):
   - If $p_{(1)} \leq \alpha_{(1)}$, reject $H_{(1)}$ and proceed to $p_{(2)}$.
   - Continue rejecting $H_{(i)}$ as long as $p_{(i)} \leq \alpha_{(i)}$.
   - Stop at the first i where $p_{(i)} > \alpha_{(i)}$; reject all null hypotheses $H_{(1)}, \ldots, H_{(i-1)}$ and do not reject the remaining hypotheses.

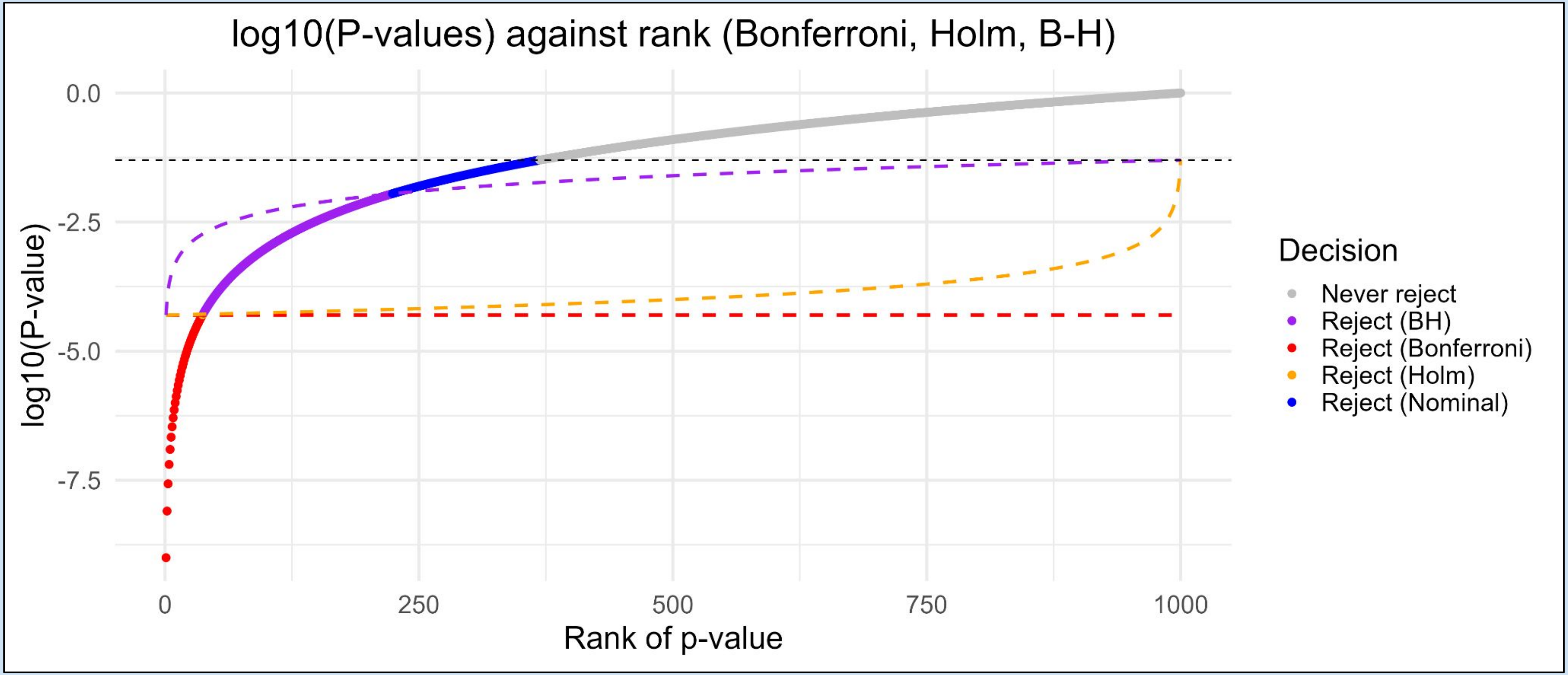
Visualising Multiplicity Corrections

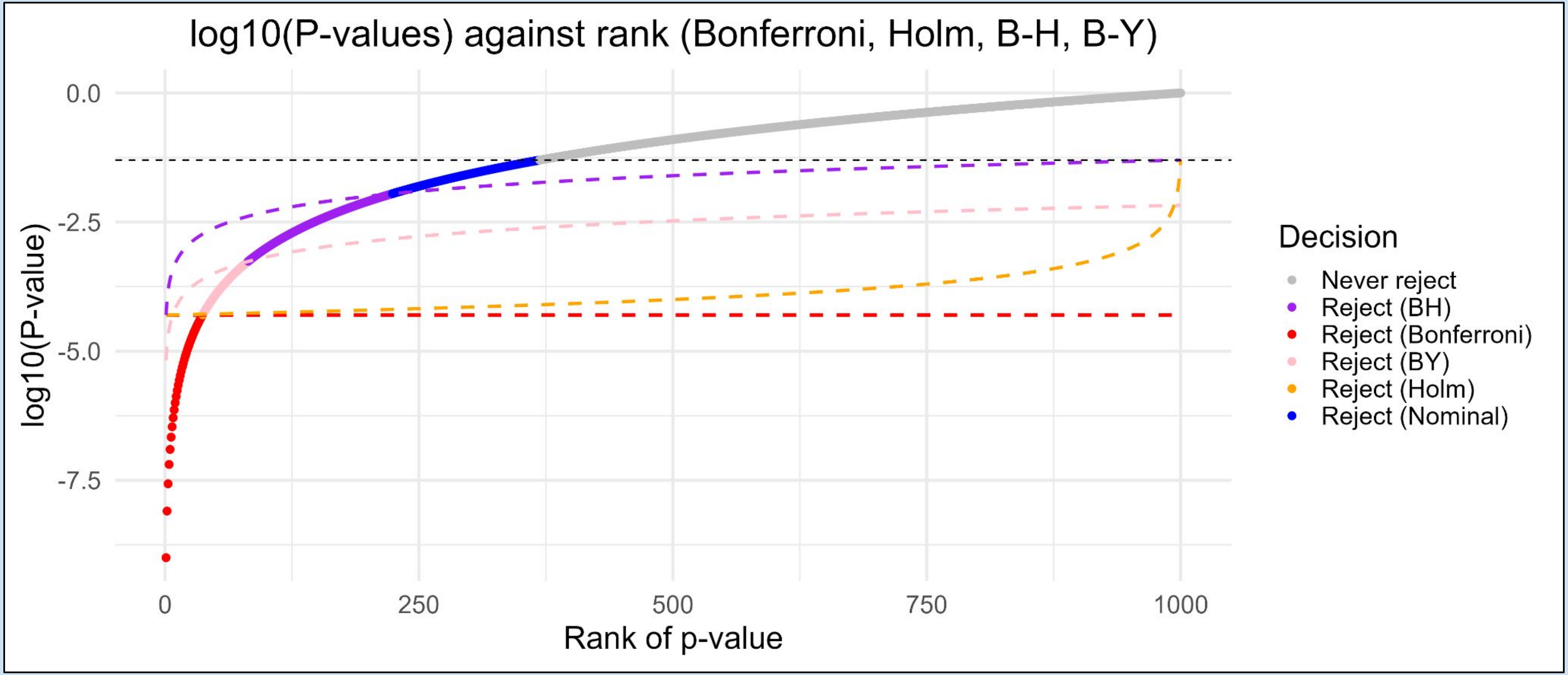

P-values against rank of p-values
P-value
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
Rank of p-value
Decision
● Do not reject (p ≥ 0.05)
● Reject (p < 0.05)


log10(P-values) against rank of p-values
log10(P-value)
0.0
-2.5
-5.0
-7.5
0
250
500
750
1000
Rank of p-value
Decision
● Do not reject ($p \geq 0.05$)
● Reject ($p < 0.05$)


log10(P-values) against rank (Bonferroni)
0.0
-2.5
-5.0
-7.5
log10(P-value)
0
250
500
750
1000
Rank of p-value
Decision
● Never reject
● Reject (Bonferroni)
● Reject (Unadjusted)


log10(P-values) against rank (Bonferroni, Holm)
log10(P-value)
0.0
-2.5
-5.0
-7.5
0
250
500
750
1000
Rank of p-value
Decision
Never reject
Reject (Bonferroni)
Reject (Holm)
Reject (Unadjusted)


log10(P-values) against rank (Bonferroni, Holm, B-H)
log10(P-value)
0.0
-2.5
-5.0
-7.5
0
250
500
750
1000
Rank of p-value
Decision
● Never reject
● Reject (BH)
● Reject (Bonferroni)
● Reject (Holm)
● Reject (Nominal)


log10(P-values) against rank (Bonferroni, Holm, B-H, B-Y)
0.0
-2.5
-5.0
-7.5
log10(P-value)
0
250
500
750
1000
Rank of p-value
Decision
● Never reject
● Reject (BH)
● Reject (Bonferroni)
● Reject (BY)
● Reject (Holm)
● Reject (Nominal)

Summary

- When performing multiple statistical tests, **corrections are necessary** in order to have useful results
- As the number of tests increases, the **family wise error rate** (probability of at least one false positive) **goes quickly to 1**
    - Bonferroni and Holm are methods to control this, **they are very conservative**
- As the **proportion of true negatives in the data increases, the false discovery rate explodes**
    - That is, the significant results will be dominated by false positives
    - B-H and B-Y control the FDR

Practical Work

Complete all the tasks in PHDS_omics_multiplicity_2025_questions.Rmd